Error analysis of regularized and unregularized least-squares regression on discretized function spaces

Dissertation

zur

Erlangung des Doktorgrades (Dr. rer. nat.)

der

Mathematisch-Naturwissenschaftlichen Fakultät

der

Rheinischen Friedrich-Wilhelms-Universität Bonn

vorgelegt von

Bastian Bohn

aus

Traben-Trarbach

Bonn 2016

Angefertigt mit Genehmigung der Mathematisch-Naturwissenschaftlichen Fakultät der Rheinischen Friedrich-Wilhelms-Universität Bonn

1. Gutachter: Prof. Dr. Michael Griebel

2. Gutachter: Prof. Dr. Jochen Garcke

Tag der Promotion: 09. Dezember 2016

Erscheinungsjahr: 2017

Zusammenfassung

In der vorliegenden Arbeit werden Regressionsverfahren zur Anpassung von Funktionen an gegebene Daten analysiert. Hierbei legen wir den Fokus auf eine Variante der Methode der kleinsten Fehlerquadrate, welche auf einer Teilmenge eines endlich-dimensionalen Vektorraums operiert.

Wir rekapitulieren in dieser Arbeit zunächst die wichtigsten Eigenschaften von Räumen vektorwertiger Funktionen. Insbesondere betrachten wir die Lebesgue-Bochner- und Sobolev-Bochner-Räume und geben eine Kurzeinführung in reproduzierende Kern-Hilberträume und reelle Interpolationsskalen. Um unsere theoretischen Resultate für die Regression immer wieder an konkreten Beispielen veranschaulichen und verdeutlichen zu können, stellen wir des Weiteren auch Fourier-Polynomräume auf hyperbolischen Kreuzen vor und geben einen Überblick über die wichtigsten Eigenschaften linearer Splineräume auf dünnen Gittern. Für Letztere liefern wir dabei auch einen Beweis für die Abschätzung der L_2-Bestapproximationsrate.

Im ersten Hauptteil der Arbeit widmen wir uns der Analyse des Regressionsproblems unter Nebenbedingungen. Hierzu stellen wir zunächst die Minimierungsaufgabe vor, welche der vektorwertigen Regression zu Grunde liegt. Wir untersuchen Existenz und Eindeutigkeit entsprechender Minima und beschäftigen uns ausführlich mit der Minimierung über beschränkten Bällen in reproduzierenden Kern-Hilberträumen. Um den Fehler eines Regressionsverfahrens über einem endlich-dimensionalen Suchraum abschätzen zu können, teilen wir diesen in zwei Summanden auf: den Diskretisierungsfehler und den Datenfehler. Wir präsentieren eine neue, auf geeigneten Jackson- und Bernstein-Ungleichungen basierende Methode, um den Diskretisierungsfehler durch den L_2 Bestapproximationsfehler zu beschränken. Des Weiteren verallgemeinern wir die existierenden Abschätzungen für den Datenfehler auf den vektorwertigen Fall. Zur Illustration unserer Resultate, ziehen wir ein Dünngitterregressionsverfahren sowie eine auf dem hyperbolischen Kreuz beruhende Methode heran. Nach geeigneter Balancierung der beiden Fehlerterme für diese Beispiele, zeigt sich, dass die zugrundeliegenden Diskretisierungen in der Lage sind, den Fluch der Dimension weitestgehend zu brechen.

Im zweiten Teil der Arbeit legen wir den Fokus auf die Regression mit Strafterm, welche dual zur Regression unter Nebenbedingungen ist. Die numerische Behandlung dieses dualen Regressionsproblems ist nun wesentlich einfacher als die entsprechende Minimierung unter Nebenbedingungen. Wir beschäftigen uns zunächst mit der Herleitung des korrespondierenden linearen Gleichungssystems und diskutieren die Vor- und Nachteile einer gitterbasierten Diskretisierung gegenüber einem kernbasierten Verfahren. Anschließend präsentieren wir eine Abschätzung für die Kondition des lin-

earen Gleichungssystems, welche für beliebige Basen des Suchraums gültig ist. Im Anschluss betrachten wir das unregularisierte Regressionsproblem unter der Annahme, dass die Eingangsdaten unverrauschte Auswertungen einer unbekannten Funktion sind. Für diesen speziellen Fall präsentieren wir eine neue obere Schranke für den zu erwartenden Fehler, welche ein deutlich besseres Abfallverhalten bezüglich der Menge der Datenpunkte aufweist als die im ersten Teil bewiesene Abschätzung. Mit Hinblick auf diesen Spezialfall untersuchen wir erneut das Verhalten des Dünngitterverfahrens und der hyperbolischen Kreuz-Methode und berechnen die optimale Kopplung zwischen dem Diskretisierungslevel und der Anzahl der Datenpunkte.

Im Anschluss zeigen wir, dass die von uns bewiesenen Abschätzungen nicht nur theoretische Relevanz haben, sondern oftmals auch die in der Praxis real zu beobachtenden Konvergenzraten widerspiegeln. Des Weiteren, stellen wir einen dimensionsadaptiven Dünngitteralgorithmus zur effektiven Behandlung anisotroper Probleme vor. Zum Abschluss präsentieren wir die Resultate, die dieser Algorithmus für Data-Mining Probleme aus realen Anwendungsgebieten erzielt.

Danksagung

Ich möchte mich an dieser Stelle bei allen bedanken, die dazu beigetragen haben, dass diese Arbeit zustande kommen konnte. Als Erstes ist hier Prof. Dr. Michael Griebel zu nennen, dem ich meinen Dank für das spannende Thema, die hervorragenden Arbeitsbedingungen am Institut für Numerische Simulation und auch für viele interessante und fruchtbare Diskussionen aussprechen möchte. Des Weiteren danke ich Prof. Dr. Jochen Garcke für zahlreiche anregende Gedankenaustausche über maschinelles Lernen sowie dafür, dass er das Zweitgutachten übernimmt.

Besonders sei auch meinen Arbeitskolleginnen und -kollegen gedankt, die mir bei Fragen und Hindernissen immer mit Rat und Tat zur Seite standen. Bei Christian Kuske, Jens Oettershagen und Markus Siebenmorgen bedanke ich mich auch für viele mathematische und nicht-mathematische Gespräche sowie das aufmerksame Korrekturlesen dieser Arbeit.

Gedankt sei hier auch Arnes, Tanja, Lisa, Felix und allen anderen Menschen, die immer wieder dazu beigetragen haben, dass ich die letzten Jahre sehr genossen habe und auch in stressigen Situationen einen kühlen Kopf bewahren konnte. Schlussendlich möchte ich auch meiner Familie einen großen Dank für Ihren Rückhalt und Ihre Unterstützung aussprechen.

Contents

1 Introduction

Data regression

In most branches of science, economy and also industry, the amount of available data has become immense during the recent years. Most of these data do not contain any valuable information at all. However, the differentiation between useful data compared to meaningless "data waste" is seldom straightforward. Prof. Dr. Johanna Wanka, Germany's Federal Minister for Education and Research, declared

> "Die Datenmengen wachsen in unserer digitalen Gesellschaft rasant.
> Wir müssen daher lernen, wie wir mit ihnen richtig umgehen können."

at the CeBIT exhibition in 2014.[1] This could be roughly translated to "The amount of data is rapidly growing in our digital society. Therefore, we have to learn how to deal with it correctly." The phenomenon of the availability of enormous amounts of data and the consequential tasks and problems arising from this are commonly summarized by the term *Big Data*.

To meet the different challenges of Big Data, such as describing the useful information in a more compact format or making predictions on future data, many ideas and approaches have emerged in the fields of machine learning, data mining and dimensionality reduction, see e.g. [16, 55, 87].

One of the most common tasks in Big Data is *regression*, i.e. the determination of a mapping $f : T \to E$ which describes the input data $(\mathbf{t}_i, \mathbf{x}_i) \in T \times E, i = 1, \ldots, n$ and allows for the prediction of the so-called data label $\mathbf{x} \in E$ of an arbitrary point $\mathbf{t} \in T$ under the assumption that $(\mathbf{t}, \mathbf{x})$ is generated by the same process that created the input data. Usually, T is an open domain in $\mathbb{R}^m$ for an $m \in \mathbb{N}$ and E is a Hilbert space. An illustrative example can be found in figure 1.1. Note that regression is also closely related to classification, see e.g. [48], and density estimation, see e.g. [40, 88].

Regression problems appear in many fields, ranging from economic time series prediction, where one looks for the prospective behavior of a time-dependent financial product, see [49], over medical causal analysis, where the correlation between specific genetic predispositions and certain clinical conditions is analyzed, see [77], to speech recognition, where audio data is parsed and interpreted, see [6]. In all of these fields, recent developments in mathematical learning theory [23, 46, 85] have led to sophisticated regression algorithms such as generalized clustering methods, radial basis function neural networks or support vector machines, see e.g. [4, 47, 50, 74].

[1] `http://www.bmbf.de/press/3580.php`

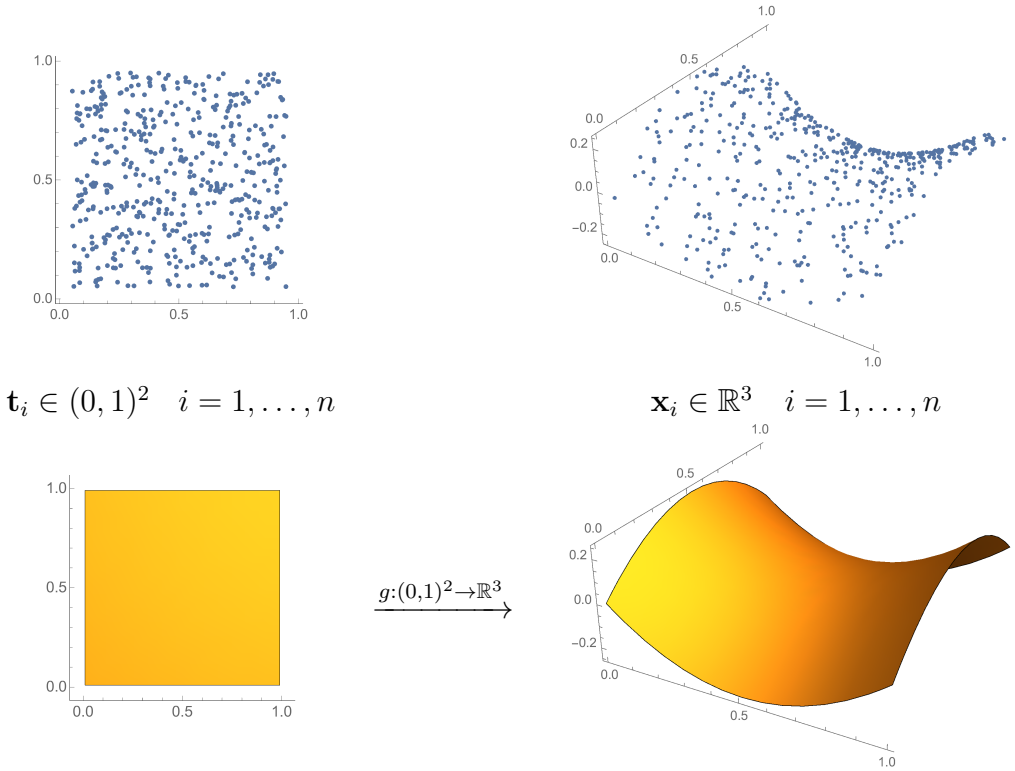

Fig. 1.1: A specific sampling, where the $\mathbf{t}_i$ are drawn uniformly in $T = (0,1)^2$ and $\mathbf{x}_i = g(\mathbf{t}_i) \in E = \mathbb{R}^3$ for a vector-valued map $g : T \to \mathbb{R}^3$. In this case the sought regressor mapping needs to be a good approximation to g.

Solving the regression problem

Given a specific method, it is desirable to make reliable estimates on its error or even on the rate of the error decay with respect to the number of data n. For certain methods, such as *least-squares* regression based on *reproducing kernel Hilbert spaces (RKHS)*, such error estimates can be derived by the approximation properties of the underlying function spaces, see [22, 75]. The main advantage of reducing the so-called *search set*, i.e. the set of candidates for f, to a subset of an RKHS is the resulting well-posedness of the underlying minimization problem for a bounded domain T and a finite-dimensional image space E. This is due to the compactness of the employed subsets in the space of continuous functions. Furthermore, estimates for bounds on the *sampling error*, which accounts for the finite sampling process, are available in terms of covering numbers. Note, however, that compactness is sufficient but not necessary for bounds on rates of convergence of the sampling error. Thus, there are also approaches based on non-compact function sets that fulfill the uniform Glivenko-Cantelli property, see e.g. [69, 89], which characterizes sets with uniform sampling error convergence. While there exist universal methods and error bounds for which the associated search set also contains non-continuous functions,

cf. [8, 9, 54], most theory and algorithms rely on minimization problems over subsets of RKHS. A thorough discussion of the error behavior in this case can be found in [23].

When dealing with an RKHS, the so-called *representer theorem*, see e.g. [58], allows to recast the regression problem into a finite system of linear equations, which can be solved in a straightforward manner. Here, the solution lies in the span of the so-called *kernel translates* $K(\mathbf{t}_i, \cdot)$, where K is the *reproducing kernel* of the search space. However, the involved computational complexity usually scales at least quadratically, or even cubically, in the amount of data n. Therefore, this approach is infeasible for problems with large input data sets, which regularly appear in Big Data applications. This is the reason why more sophisticated methods have to be developed in order to cope with this non-beneficial scaling in n, see e.g. [74]. Besides the problem of large sample sizes, there is also another difficulty in applying the representer theorem: If we do not have access to a closed form of the corresponding kernel function K, e.g. for infinite series kernels, a discretized version of the kernel must be used for solving the regression problem. Then, we naturally encounter the task of determining a suitable discretization accuracy for the kernel function. Usually, this requires a careful consideration of the kernel structure at hand. We refer the interested reader to [43, 44], where the truncation of series kernels is considered which stem either from tight frames in Hilbert spaces or, in the more general setting, from multiscale expansions in certain Besov spaces.

One way to circumvent the above-mentioned problems is to consider a discretization of the search space where the position and the shape of a basis function are independent of the data points in contrast to the basis $K(\mathbf{t}_i, \cdot)$. Probably the most common examples for such a discretization are ansatz spaces based on (tensor product) grids. Here, the size of the resulting system of equations does no longer depend on the amount of data n, but scales directly with the grid level k and, thus, the size of the grid. It can be shown straightforwardly that the computational costs of such an approach scale only linearly in n, see e.g. [31, 35]. Since we can no longer rely on the representer theorem, the usage of such a discretization directly implies an additional error which has to be controlled. This is in general not an easy task, especially for norm-regularized regression problems.

Despite getting rid of the computational complexity problem with respect to n, there is still a limiting factor when considering the performance of regression algorithms on a grid space: the dimension m of the domain T. This is due to the fact that, for standard tensor product grids, the degrees of freedom N_k of a grid space V_k have to scale exponentially in m, e.g. $N_k \simeq 2^{km}$, to achieve (roughly) the same approximation error as in the case of univariate regression ($m = 1$) and 2^k degrees of freedom. This phenomenon is known as the so-called *curse of dimensionality*, see [5]. It prevents us from applying full tensor-product grid approaches in multivariate settings with $m > 3$.

However, if the problem at hand fulfills certain additional regularity assumptions, such as mixed Sobolev smoothness of the solution for instance, the curse of dimensionality can be broken to some extent by using sparse ansatz spaces based on e.g. *sparse grids* or *hyperbolic crosses*, see [14, 78]. For these spaces, the curse of dimensionality only appears in the logarithm of the basis size, i.e. $N_k \simeq 2^k k^{m-1}$ guarantees the same discretization

error - up to logarithms - as taking 2^k degrees of freedom in the univariate case. This allows us to use sparse grids and hyperbolic crosses to interpolate and approximate functions on domains with dimension $m > 3$. Naturally, the question arises if this beneficial behavior in approximation problems carries over to sparse grid and hyperbolic cross regression algorithms as introduced in [11, 35, 68].

Error analysis

The main goal of this thesis is to provide an error analysis of the regression problem over finite-dimensional search spaces such as sparse grid or hyperbolic cross spaces. Furthermore, we aim to show that an appropriate coupling between the discretization scale k and the number of data n leads to convergence results for which the curse of dimensionality is indeed broken for sparse ansatz spaces if certain smoothness conditions are met. To this end, we investigate two different approaches to analyze the regression problem over finite-dimensional search spaces.

First, we consider the behavior of the *constrained* least-squares regression error in a very general setting. Here, each data point $(\mathbf{t}_i, \mathbf{x}_i)$ is assumed to be drawn according to some compactly supported measure ρ. The regressor function f from the search set $V_{k,b}$, which is a centered ball of radius b in the finite-dimensional search space V_k, is now the solution of the minimization problem

$$f := \operatorname*{arg\,min}_{h \in V_{k,b}} \frac{1}{n} \sum_{i=1}^{n} \|h(\mathbf{t}_i) - \mathbf{x}_i\|_E^2.$$

Thus, a minimizer f not only needs to be a good fit for the data, but it also has to be smooth enough to have a V_k norm smaller than the parameter b. Based on the additive splitting of the overall regression error into a *bias* part and a *sampling error* part, as it is done in [22, 23, 75, 79], each part of the error can be considered separately. When dealing with finite-dimensional search spaces, the bias is also called *discretization error*. However, the techniques which are commonly used to deal with the bias in the infinite-dimensional case, see [23], cannot simply be applied to the discretization error. Although there exist first estimates on the discretization error in dependence on k in special cases, see e.g. [36, 53, 61, 92], there is not yet a result which is applicable for a broader choice of search spaces V_k.

In a second step, we focus on a *penalized* regression problem, which can be shown to be dual to the constrained problem. Furthermore, we have a closer look at noiseless function regression in the limit case of *unregularized* regression, where the penalty term vanishes. Here, the data points are assumed to be evaluations of an unknown function $g : T \to E$, i.e. $(\mathbf{t}_i, \mathbf{x}_i) = (\mathbf{t}_i, g(\mathbf{t}_i))$ for each $i = 1, \ldots, n$. In unregularized regression, the search set equals the search space and the corresponding minimization problem becomes

$$f := \operatorname*{arg\,min}_{h \in V_k} \frac{1}{n} \sum_{i=1}^{n} \|h(\mathbf{t}_i) - g(\mathbf{t}_i)\|_E^2.$$

Here, it is possible to achieve higher-order convergence rates than for the more general regularized problem above. First results for regression with orthonormal basis sets of V_k can be found in [19, 21, 60, 61]. However, a similar result for arbitrary bases is not available yet.

Contributions of this thesis

In this thesis, we provide upper bounds on the overall error for both regularized and unregularized regression and apply them to finite-dimensional grid spaces. For a constrained regression method based on sparse grids, for instance, we show that the error with respect to the number of sample points n is bounded by

$$\mathcal{O}\left(n^{-\frac{2s}{2s+1}} \log(n)^m\right)$$

under certain conditions. Here, $0 < s \leq 2$ denotes the Sobolev degree of mixed smoothness of the space in which the true solution resides. In the unregularized case, we even obtain the bound

$$\mathcal{O}\left(n^{-2s} \log(n)^{(2s+1)m-1}\right)$$

for noiseless function regression. Both of these results are the first of their kind for sparse grids. Our own contributions in the context of regression error estimates and convergence analysis can be summarized as follows:

- For the constrained problem, we provide a coupling between the search set radius b, which controls the regularization, and the number of degrees of freedom N_k of a finite-dimensional search space V_k such that the discretization error is proportional to the L_2 best approximation error in V_k. With the help of this result, we present error bounds for the sparse grid-based and the hyperbolic cross-based constrained regression methods and their full grid counterparts.

- We extend the analysis of the unregularized problem from [21] to the penalized problem and to arbitrary basis sets of V_k in order to obtain a stability result which ensures the numerical solvability of the corresponding system of equations in dependence on the regularization parameter and the condition number of the stiffness matrix.

- With the help of the above-mentioned stability result, we derive an upper bound on the error of the unregularized regression problem for arbitrary basis sets. Our results are applied to full grids, sparse grids and hyperbolic crosses to analyze the behavior of these spaces for noiseless function regression.

- Since the discretization error and the sampling error need to be balanced to obtain optimal convergence rates in terms of the number of samples n, we derive an appropriate coupling between k and n to achieve these optimal rates.

- We complement our theoretical analysis by numerical experiments and observe the behavior of a dimension-adaptive sparse grid algorithm in practical applications. To this end, we enhanced the C++ *sparselib* code developed in [30] to (adaptively) solve the (un)penalized regression problem.

Outline

The remainder of this thesis is organized as follows: In chapter 2, we introduce some notations which we frequently use. Furthermore, we briefly comment on some notational peculiarities which appear in the course of this thesis, e.g. our formal restriction to real vector spaces.

As certain vector-valued function spaces appear throughout the thesis, we give a proper definition of them in chapter 3. Here, we start with the basics on Lebesgue and Sobolev spaces before considering the concept of reproducing kernel Hilbert spaces and real interpolation scales. Furthermore, we also provide the necessary details on the finite-dimensional grid spaces which serve as examples in the later chapters.

In chapter 4, we present the constrained regression problem in a very general setting and hint at its basic properties. We provide proofs on the existence and uniqueness of its solutions under specific conditions. After introducing the bias/sampling error splitting, we derive upper bounds on each part of the error and apply them to full grid, sparse grid and hyperbolic cross examples. We also present a comparison of our results to the ones of other researchers.

The penalized regression problem, which stems from the Lagrangian dual formulation of the constrained problem, is investigated in chapter 5. After relating the primal and the dual formulation, we discuss how to recast the regression problem into a system of linear equations in order to solve it algorithmically. We derive a probabilistic stability result which paves the way for our convergence analysis in the case of unregularized, noiseless regression. Again, we consider how these results can be applied to our example settings and discuss how they relate to recent research.

To support our theoretical results, we provide numerical examples on the behavior of the regression error in chapter 6. Besides investigating convergence rates for smooth toy problems, we also discuss an adaptive sparse grid algorithm, which reduces the computational costs when dealing with anisotropic regressor functions significantly compared to the standard approach. Furthermore, we have a look at the performance of the adaptive algorithm for real world problems.

Finally, we conclude this thesis in chapter 7 by giving a summary and an outlook on the next steps regarding the analysis of finite-dimensional regression algorithms.

2 Notation

We now give a few details on the notation we use throughout this thesis and hint at peculiarities in this context. An overview of the most important variables, functions, spaces, etc. which we deal with when analyzing the regression problem can be found in the introduction of chapters 4 and 5.

Vectors, multiindices and norms

We use the bold face notations $\mathbf{l} \in \mathbb{N}^m, \mathbf{x} \in \mathbb{R}^m$ for multidimensional indices or vectors. Unless stated otherwise, the elements of an index or vector are then denoted by $l_1, \dots, l_m \in \mathbb{N}$ and $x_1, \dots, x_m \in \mathbb{R}$ (analogously for different letters). This must not be confused with the bold-face notation for a collection of vectors $\mathbf{x}_1, \dots, \mathbf{x}_n$ for instance. Sometimes, we also use the arrow-notation $\vec{\alpha}$ to denote a vector. This is of special importance when dealing with vectors of vectors, e.g. $\vec{\mathbf{x}} = (\mathbf{x}_1, \dots, \mathbf{x}_n)^T$. Inequalities of type $\mathbf{l} \leq \mathbf{k}$ or $\mathbf{l} \geq \mathbf{k}$ have to be understood componentwise. The vector $\mathbf{e}_j \in \mathbb{R}^m$ denotes the j-th unit vector, i.e. $e_j = 1$ and $e_k = 0$ for all $k \in \{1, \dots, m\} \setminus \{j\}$.

The expression $|a|$ stands for the absolute value of a scalar $a \in \mathbb{R}, \mathbb{C}$ or $\mathbb{Z}$. Analogously, we write $|A|$ for the cardinality of a set A. The ℓ_p vector and index norms are denoted by $\|\mathbf{x}\|_{\ell_p} := (\sum_{i=1}^m x_i^p)^{\frac{1}{p}}$ or simply by $\|\mathbf{x}\|_p$ for $1 \leq p < \infty$ and $\|\mathbf{x}\|_{\ell_\infty} := \max_{i=1,\dots,m} |\mathbf{x}_i|$ or $\|x\|_\infty$ for $p = \infty$.

Asymptotics and constants

Although we only present asymptotic rates in our main results, the behavior of the involved constants can often be derived easily by a detailed investigation of the proofs which we provide. Throughout this thesis, we use the Landau symbol $f(a) = \mathcal{O}(g(a))$ for $a \to \infty$ to denote that there exist constants $a_0, c > 0$ such that $f(a) \leq cg(a)$ for all $a > a_0$. Similarly, we use $f(a) \lesssim g(a)$ to denote that there exists a $c > 0$ such that $f(a) \leq cg(a)$ for all a. Furthermore, $f(a) \simeq g(a)$ means $f(a) \lesssim g(a)$ and $g(a) \lesssim f(a)$. Finally, if the functions f and g depend on different parameters a and b, we write $f(a) \lesssim g(b)$ to denote that $a := a(b)$ is coupled to b such that $f(a(b)) \lesssim g(b)$. Furthermore, we write $f(a) \ll g(b)$ if the coupling between a and b fulfills $a(b) \to \infty$ and $f(a(b)) = o(g(b))$ for $b \to \infty$. Here, the little-o Landau symbol is used, i.e. for all $c > 0$ there exists a $b_0 > 0$ such that $f(a(b)) \leq cg(b)$ for all $b > b_0$.

Notational ambiguousness

We try to avoid ambiguous notation wherever possible. However, some characters can have a different meaning depending on the context they appear in. Take $\mathcal{L}$ for instance: Throughout this thesis, we use $\Sigma_{\mathcal{L}}$ to denote the Lebesgue σ-algebra. Furthermore, $\mathcal{L}(A; B)$ denotes the space of bounded linear operators from A to B and $\mathcal{L}_{\mathcal{Z}_n,b}$ denotes the Lagrangian functional in chapter 5. Even though the character $\mathcal{L}$ is used multiple times here, it is directly clear from the context how it has to be understood.

Variable dependencies

For the ease of notation, we sometimes use variable names which do not automatically reveal on what parameters the variable might depend, such as the cost function bound M_ψ introduced in section 4.4, which might depend on the search set radius b. Another example is the mass matrix M of the space V_k. It should be clear that it depends on the specific choice of the basis $\nu_1, \dots, \nu_{N_k}$. Hence, we use M as a short notation for $M(\nu_1, \dots, \nu_{N_k})$.

In the case of the bound $K(N_k)$ introduced in chapter 5, we employ this specific notation to be consistent with the notation from [19, 21, 60]. A more precise notation in this case would be $K(\nu_1, \dots, \nu_{N_k})$ since the size of this bound does not only depend on the number of basis functions but also on the specific choice of the basis in our case.

Complex image spaces $E = \mathbb{C}^d$

Throughout this thesis we consider vector-valued mappings $f : T \to E$, where $T \subset \mathbb{R}^m$ is a bounded, open domain and $(E, \langle \cdot, \cdot \rangle_E)$ is a Hilbert space. Although all of our results in chapters 4 and 5 are stated only for the case where E is a vector space over $\mathbb{R}$, the corresponding variants for the complex Hilbert spaces $E = \mathbb{C}^d, d \in \mathbb{N}$, can usually be derived analogously by using the isomorphy $\mathbb{C} \cong \mathbb{R}^2$ and changing the scalar product of E accordingly. For the sake of brevity and readability, we omit the complex versions throughout the thesis. Nevertheless, when dealing with $E = \mathbb{C}^d$ for the study of Fourier polynomials and when referring to our results from the corresponding chapters, we implicitly mean the complex version of the corresponding result. We refer the interested reader to [66] for a thorough discussion of vector-valued reproducing kernel Hilbert spaces in the case of complex E. A detailed explanation on how to derive the matrix vector equations for the dual regression problem which we deal with in chapter 5 can be found in [62] for the case of complex-valued functions.

The terms “constrained”, “penalized” and “regularized”

Since we often use the terms *(un)constrained*, *(un)penalized* or *(un)regularized* regression, let us shortly clarify what we mean by that. By “constrained regression”, we refer to a problem where the search set, i.e. the set of all functions which are considered in the

corresponding minimization problem, implicitly carries a norm constraint. This means that only functions f which fulfill $\|f\| \leq b$ for some norm $\|\cdot\|$ and some $0 < b < \infty$ are taken into account. The term "penalized" refers to a problem, where the norm bound is not forced onto the search set, but a penalty term $\mu\|f\|$ with $\mu > 0$ is added to the minimization functional to obtain a minimizer with small norm. Under certain assumptions, the constrained and the penalized regression problem are dual to each other, see section 5.1. This is also the reason why the constrained problem is often referred to as the *primal problem* and the penalized problem is called the *dual problem*. "Regularized" can mean either the constrained problem or the penalized problem. By "regularized" we just mean that the original regression problem is altered in such a way that the existence of a solution is guaranteed. This also clarifies that the term "regularization parameter" refers to the norm bound b when talking about the constrained problem and it refers to the (Lagrange) parameter μ when we investigate the penalized problem. Note, finally, that "unconstrained", "unpenalized" and "unregularized" refer to the same problem, where no regularization is done, i.e. there is no constraint and no penalty term in the problem formulation.

Search spaces $\mathcal{H}$ and V_k

When considering the search space of a regression problem, we use either $\mathcal{H}$ or V_k, depending on the specific situation. If not explicitly stated otherwise, $\mathcal{H}$ can refer to either an infinite- or a finite-dimensional space. However, V_k is always finite-dimensional. Therefore, if we want to emphasize that we are working in a finite-dimensional setting or if we deal with a whole scale of search spaces, we will always use the notation V_k for $k \in \mathbb{N}$. If the dimension of the search space is irrelevant, we use $\mathcal{H}$.

3 Function spaces

In this chapter we will briefly recapitulate the most important function spaces which are relevant for the remainder of this thesis. Note that this is by no means a complete and profound analysis but rather serves to give a short overview and illustrate important properties of these spaces.

First, we introduce the Lebesgue spaces $L_{p,\omega}$ for generic positive measures ω and $1 \leq p \leq \infty$ in section 3.1. Furthermore, we generalize this concept to vector-valued functions and consider the corresponding Bochner spaces. Then, we have a look at the Bessel potential Sobolev spaces H_p^s with respect to the Lebesgue measure and with $1 \leq p \leq \infty$ and $s \geq 0$ in section 3.2. Also in this case, we consider the vector-valued variants of these spaces. In section 3.3, we present a short introduction to vector-valued reproducing kernel Hilbert spaces and their basic properties. Subsequently, we present the $\mathbb{K}$-functional and real interpolation scales in section 3.4. Finally, we consider finite-dimensional discretization spaces based on full grids, sparse grids and hyperbolic crosses in section 3.5.

3.1 Lebesgue spaces

The definitions and concepts introduced here are based on [1, 3]. We assume that the reader is already familiar with the notion of σ-algebras and (positive) measures thereon. For a formal definition of generic σ-algebras and measures we refer to [1, 3].

Definition 3.1 [Lebesgue σ-algebra and Lebesgue measure]
Let the σ-algebra $\Sigma_{\mathcal{L}} = \Sigma_{\mathcal{L}}(\mathbb{R}^m)$ over $\mathbb{R}^m$ and the measure λ over $\Sigma_{\mathcal{L}}$ have the following properties:

- *If $X \subset \mathbb{R}^m$ is open, then $X \in \Sigma_{\mathcal{L}}$.*
- *For each $Y \in \Sigma_{\mathcal{L}}$ with $\lambda(Y) = 0$, it holds that each subset $X \subset Y$ is an element of $\Sigma_{\mathcal{L}}$ and $\lambda(X) = 0$.*
- *For the hyperrectangle $R := \prod_{i=1}^m [a_i, b_i]$ with $b_i \geq a_i$ for all $i = 1, \ldots, m$, it holds $R \in \Sigma_{\mathcal{L}}$ and $\lambda(R) = \prod_{i=1}^m (b_i - a_i)$.*
- *Let $X \in \Sigma_{\mathcal{L}}$. Then for each $\mathbf{t} \in \mathbb{R}^m$ we have $X + \mathbf{t} := \{\mathbf{s} + \mathbf{t} \mid \mathbf{s} \in X\} \in \Sigma_{\mathcal{L}}$ and $\lambda(X + \mathbf{t}) = \lambda(X)$.*

*We call $\Sigma_{\mathcal{L}}$ **Lebesgue σ-algebra** and λ **Lebesgue measure** on $\mathbb{R}^m$.*

Note that we determined both $\Sigma_{\mathcal{L}}$ and λ by the above definition. Another important σ-algebra is the Borel algebra.

Definition 3.2 [Borel σ-algebra]
Let S be a separable normed space. The smallest σ-algebra $\Sigma_{\mathcal{B}} = \Sigma_{\mathcal{B}}(S)$ which contains all open subsets of S is called ***Borel σ-algebra****.*

In the following, we consider the pair (S, Σ) of a separable normed space S and a σ-algebra Σ on it. We assume that either $\Sigma = \Sigma_{\mathcal{B}}$ holds or $S = \mathbb{R}^m$ and $\Sigma = \Sigma_{\mathcal{L}}$ hold. However, note that most of the considerations below are valid also for other choices of S and Σ.

3.1.1 Lebesgue spaces of scalar-valued functions

First we introduce the concept of measurability.

Definition 3.3 [Measurable Function]
Let $T \in \Sigma$. A function $f : T \to [-\infty, \infty]$ is called ***measurable*** *if*

$$\{\mathbf{t} \in T \mid f(\mathbf{t}) > a\} \in \Sigma \quad \forall\, a \in \mathbb{R}.$$

It is easy to see that a so-called simple function

$$s(\mathbf{t}) = \sum_{j=1}^{N} a_j \chi_{T_j}(\mathbf{t}) \tag{3.1}$$

is measurable for arbitrary $N \in \mathbb{N}$ with $a_j \in \mathbb{R}$, $T_j \in \Sigma$ and

$$\chi_{T_j}(\mathbf{t}) := \begin{cases} 1 & \text{if } \mathbf{t} \in T_j \\ 0 & \text{else} \end{cases}$$

for all $j = 1, \ldots, N$. Furthermore, it can be shown that, for each measurable function $f : T \to \mathbb{R}$, there exists a sequence of simple functions $f_i : T \to \mathbb{R}$ which converge to f pointwise. If f is non-negative, the f_i can be chosen monotonically increasing in each point $\mathbf{t} \in T$. With these results at hand, we can finally define the Lebesgue integral of a measurable function.

Definition 3.4 [Lebesgue integral]
Let $T \in \Sigma$ and let $(T, \Sigma(T), \omega)$ be a measure space, where $\Sigma(T) := \{X \in \Sigma \mid X \subseteq T\}$. Let $a_j \geq 0$ and $T_j \subset T$ for all $j = 1, \ldots, N$ for the simple function s in (3.1)*. We define its* ***Lebesgue integral*** *by*

$$\int_T s \, \mathrm{d}\omega := \sum_{j=1}^{N} a_j \omega(T_j),$$

where we set $0 \cdot \infty = 0$. Now let $f : T \to [0, \infty)$ be measurable with respect to $\Sigma(T)$ and

let $(f_i)_{i=1}^{\infty}$ be a monotonically (in each point) increasing sequence of non-negative simple functions which converges to f pointwise. Then, the Lebesgue integral of f is defined by

$$\int_T f \,\mathrm{d}\omega := \lim_{i\to\infty} \int_T f_i \,\mathrm{d}\omega.$$

For a measurable function $f : T \to \mathbb{R}$, we define the Lebesgue integral by

$$\int_T f \,\mathrm{d}\omega := \int_T \max(f, 0) \,\mathrm{d}\omega - \int_T \max(-f, 0) \,\mathrm{d}\omega$$

if at least one of the terms on the right hand side is finite.

Definition 3.5 [Lebesgue spaces $L_{p,\omega}(T)$]
Let $(T, \Sigma(T), \omega)$ be as above and let $1 \le p \le \infty$. The space

$$L_{p,\omega}(T) := \{f : T \to \mathbb{R} \text{ measurable} \mid \|f\|_{L_{p,\omega}(T)} < \infty\}$$

with

$$\|f\|_{L_{p,\omega}(T)} := \left(\int_T |f|^p \,\mathrm{d}\omega\right)^{\frac{1}{p}}$$

if $1 \le p < \infty$ and

$$\|f\|_{L_{\infty,\omega}(T)} := \operatorname*{ess\,sup}_{\mathbf{t}\in T} |f(\mathbf{t})| = \inf\{a \ge 0 \mid \omega(\{\mathbf{t} \in T \mid |f(\mathbf{t})| > a\}) = 0\}$$

*if $p = \infty$ is called **Lebesgue space**. If $\omega = \lambda_T := \lambda|_T$, i.e. we deal with the restriction of the Lebesgue measure to T, we write $L_p(T)$ for $L_{p,\lambda_T}(T)$. We often also write $\int_T f(\mathbf{t}) \,\mathrm{d}\mathbf{t}$ instead of $\int_T f \,\mathrm{d}\lambda_T$ in this case.*

Note that functions which only differ on a nullset, i.e. a set $A \in \Sigma(T)$ such that $\omega(A) = 0$, have the same $L_{p,\omega}(T)$ norm. Therefore, we identify such functions with each other. In a strict sense, the elements of $L_{p,\omega}(T)$ are only equivalence classes with respect to this identification. However, in this thesis, we ignore this formal detail when talking about elements of the Lebesgue spaces and consider functions instead of equivalence classes.

The Lebesgue spaces $L_{p,\omega}(T)$ are Banach spaces for all $1 \le p \le \infty$. Additionally, $L_{2,\omega}(T)$ is a Hilbert space with the inner product

$$\langle f, g\rangle_{L_{2,\omega}(T)} = \int_T f \cdot g \,\mathrm{d}\omega.$$

Two very important properties of the Lebesgue integral are the inequality

$$\left|\int_T f \,\mathrm{d}\omega\right| \le \int_T |f| \,\mathrm{d}\omega. \tag{3.2}$$

and the Hölder inequality

$$\int_T |f \cdot g| \, \mathrm{d}\omega \leq \|f\|_{L_{p,\omega}(T)} \|g\|_{L_{q,\omega}(T)}, \tag{3.3}$$

which holds for all Lebesgue measurable f, g and for all $1 \leq p, q \leq \infty$ with $p^{-1} + q^{-1} = 1$. In the case $p = q = 2$, this becomes the Cauchy-Schwarz inequality for Hilbert spaces.

Since we are mostly dealing with probability measures on $T \subset S$ in this thesis, we shortly examine the special case $\omega(T) = 1$. Here, the inequality (3.2) can be generalized to Jensen's inequality, which states

$$g\left(\int_T f \, \mathrm{d}\omega\right) \leq \int_T g \circ f \, \mathrm{d}\omega \tag{3.4}$$

for each $g : \mathbb{R} \to \mathbb{R}$ which is convex on the image of f. From this, one obtains

$$\|f\|_{L_{p,\omega}(T)} \leq \|f\|_{L_{q,\omega}(T)} \tag{3.5}$$

for every $f \in L_{q,\omega}(T)$ and $1 \leq p \leq q \leq \infty$, which shows that $L_{q,\omega}(T)$ is continuously embedded into $L_{p,\omega}(T)$.

3.1.2 Lebesgue-Bochner spaces of vector-valued functions

In this subsection, we deal with an extension of the concepts of subsection 3.1.1 to vector-valued function spaces. For a thorough consideration of Bochner integrability and Bochner integrals, we refer to [3], on which our review is based. As in subsection 3.1.1, let S be a separable normed space and let $\Sigma \in \{\Sigma_{\mathcal{B}}, \Sigma_{\mathcal{L}}\}$, where $S = \mathbb{R}^m$ if $\Sigma = \Sigma_{\mathcal{L}}$. For $T \in \Sigma$, we let ω be a measure defined on the restriction $\Sigma(T)$. Furthermore, let $(E, \|\cdot\|_E)$ be a separable Hilbert space[1] for which the norm $\|\cdot\|_E$ is induced by the scalar product $\langle \cdot, \cdot \rangle_E$. We define measurability analogously to the scalar-valued case.

Definition 3.6 [Lebesgue measurable function]
*A function $f : T \to E$ is called **(Bochner) measurable** if $f^{-1}(A) \in \Sigma(T)$ for every open set $A \subseteq E$.*

It can be shown that this implies that there exists a sequence $(f_i)_{i=1}^{\infty}$ of E-valued simple functions which converge to f pointwise almost everywhere. To this end, let

$$f_i(\mathbf{t}) := \sum_{j=1}^{N_i} a_{i,j} \chi_{T_{i,j}}(\mathbf{t})$$

for $i \in \mathbb{N}$, where $N_i \in \mathbb{N}, a_{i,j} \in E$ and $T_{i,j} \in \Sigma(T)$. If these simple functions can be

[1] Note that the whole Lebesgue-Bochner theory also holds for separable Banach spaces instead of Hilbert spaces. However, when considering Sobolev-Bochner spaces and reproducing kernel Hilbert spaces later on, it makes sense to restrict ourselves to the case where E is a Hilbert space.

chosen such that $\omega(T_{i,j}) < \infty$ whenever $a_{i,j} \neq 0$, we call them *integrable*.

Definition 3.7 [Bochner integrability and Bochner integral]
Let $f : T \to E$ be measurable and let there exist a sequence of integrable step functions $(f_i)_{i=1}^{\infty}$ which converge to f pointwise almost everywhere. If

$$\lim_{i\to\infty} \int_T \|f - f_i\|_E \, \mathrm{d}\omega = 0,$$

where the integral has to be understood in the Lebesgue sense, we call f ***(Bochner) integrable****. The corresponding* ***Bochner integral*** *of f is defined by*

$$\int_T f \, \mathrm{d}\omega = \lim_{i\to\infty} \sum_{j=1}^{N_i} a_{i,j}\omega(T_{i,j}).$$

It can easily be verified that the integral operator $f \to \int_T f \, \mathrm{d}\omega$ is linear. Due to the definition of Bochner integrability, it can be shown that $f : T \to E$ is Bochner integrable with respect to ω if and only if $\|f(\cdot)\|_E : T \to \mathbb{R}$ is Lebesgue integrable with respect to ω. This leads to the definition of the (Lebesgue-)Bochner spaces, which are the vector-valued analogue to the scalar-valued Lebesgue spaces $L_{p,\omega}(T)$ from subsection 3.1.1.

Definition 3.8 [(Lebesgue-)Bochner space]
Let $T \in \Sigma$ and let $1 \leq p \leq \infty$ be fixed. We define the ***(Lebesgue-)Bochner space*** *$L_{p,\omega}(T; E)$ with respect to the measure ω by*

$$L_{p,\omega}(T; E) := \left\{ f : T \to E \text{ Bochner integrable } \mid \|f\|_{L_{p,\omega}(T;E)} < \infty \right\}.$$

Here, the Bochner norm is defined as

$$\|f\|_{L_{p,\omega}(T;E)} := \left(\int_T \|f(\mathbf{t})\|_E^p \, \mathrm{d}\omega(\mathbf{t}) \right)^{\frac{1}{p}}$$

if $1 \leq p < \infty$ and as

$$\|f\|_{L_{\infty,\omega}(T;E)} := \operatorname*{ess\,sup}_{\mathbf{t}\in T} \|f(\mathbf{t})\|_E = \inf \{a \geq 0 \mid \omega(\{\mathbf{t} \in T \mid \|f(\mathbf{t})\|_E > a\}) = 0\}$$

if $p = \infty$. Analogously to the scalar-valued case, we write $L_p(T; E)$ for $L_{p,\lambda_T}(T; E)$.

As in the scalar-valued case, all Bochner spaces are Banach spaces and $L_{2,\omega}(T; E)$ is a Hilbert space since E is also a Hilbert space. Let us shortly review the most important properties which carry over from Lebesgue integrals to Bochner integrals. Similar to (3.2), we have

$$\left\| \int_T f \, \mathrm{d}\omega \right\|_E \leq \int_T \|f\|_E \, \mathrm{d}\omega \tag{3.6}$$

also in the Bochner case. Secondly, for any continuous linear operator $P : E \to F$ which maps into a separable Hilbert space F, we obtain

$$P\left(\int_T f \,\mathrm{d}\omega\right) = \int_T P(f) \,\mathrm{d}\omega \tag{3.7}$$

which is analogous to the linearity of the integral operator in the scalar-valued case, where this observation is trivial. If ω is a probability measure, i.e. $\omega(T) = 1$, we obtain

$$\|f\|_{L_{p,\omega}(T;E)} \leq \|f\|_{L_{q,\omega}(T;E)} \tag{3.8}$$

for $f \in L_{q,\omega}(T)$ and $1 \leq p \leq q \leq \infty$ in analogy to (3.5).

3.2 Sobolev spaces

For Sobolev spaces, we restrict our short recapitulation to the measure space $(\mathbb{R}^m, \Sigma_{\mathcal{L}}, \lambda)$, where λ is the Lebesgue measure on $\mathbb{R}^m$. Therefore, in the notation of section 3.1, we have $S = \mathbb{R}^m$, $\Sigma = \Sigma_{\mathcal{L}}$ and $\omega = \lambda$. Our review of definitions and results for isotropic Sobolev spaces is based on [1, 7, 80]. For the case of Sobolev spaces of mixed smoothness, we refer to [52, 76, 84]. The special case of Sobolev-Bochner spaces of vector-valued functions follows [72, 76, 81].

Since we consider the Fourier transform to define the Bessel potential Sobolev spaces, we need to deal also with complex-valued functions. For ease of notation, we write $L_p(T)$ for the Lebesgue space on the domain T regardless of whether we deal with real-valued or complex-valued functions.[2]

3.2.1 Sobolev spaces of scalar-valued functions

In this thesis, we are mainly interested in the so-called Bessel potential Sobolev spaces $H_p^s(T) \subset L_2(T)$ for $s > 0$ defined on an open domain $T \in \Sigma_{\mathcal{L}}$ with Lipschitz boundary.

We now briefly review the definition of tempered distributions and the concepts of the Fourier transform and weak derivatives. We refer the interested reader to [1, 7, 80] for details. For a sufficiently smooth complex-valued $f : \mathbb{R}^m \to \mathbb{C}$, let

$$D^{\mathbf{i}} f = \frac{\partial^{|\mathbf{i}|_{\ell_1}}}{\partial_{t_1}^{i_1} \cdots \partial_{t_m}^{i_m}}$$

denote the differentiation operator with respect to the multiindex $\mathbf{i} \in \mathbb{N}^m$.

[2] Formally, one has to consider the space $L_p(T;\mathbb{C}) \cong L_p(T;\mathbb{R}^2)$ for complex-valued functions.

Definition 3.9 [Schwartz space and tempered distributions]
The space of all infinitely differentiable functions $f : \mathbb{R}^m \to \mathbb{C}$ for which

$$\sup_{\mathbf{t}\in\mathbb{R}^m} (1 + \|\mathbf{t}\|_2)^k |D^{\mathbf{i}} f(\mathbf{t})| \tag{3.9}$$

*is finite for all $k \in \mathbb{N}$ and $\mathbf{i} \in \mathbb{N}^m$ is called the **Schwartz space** $\mathcal{S} := \mathcal{S}(\mathbb{R}^m)$. The topology on $\mathcal{S}$ is defined by the seminorms (3.9). Its continuous dual space $\mathcal{S}' := \mathcal{S}'(\mathbb{R}^m)$ is called the space of **tempered distributions**.*

It can be shown that, for each $f \in L_p(\mathbb{R}^m)$, the functional

$$g \in \mathcal{S} \to \int_{\mathbb{R}^m} f \cdot g \, \mathrm{d}\lambda$$

is a tempered distribution and, in this sense, we obtain $L_p(\mathbb{R}^m) \subset \mathcal{S}'(\mathbb{R}^m)$ for every $1 \leq p \leq \infty$.

Definition 3.10 [Weak derivatives]
*The **weak derivative** $D^{\mathbf{i}} f \in \mathcal{S}'$ of $f \in \mathcal{S}'$ is given by*

$$D^{\mathbf{i}} f(g) := f\left(D^{\mathbf{i}} g\right) \tag{3.10}$$

for any $g \in \mathcal{S}$. For $f \in L_p(\mathbb{R}^m)$ and $D^{\mathbf{i}} f \in L_p(\mathbb{R}^m)$, this leads to the common notation

$$\int_{\mathbb{R}^m} D^{\mathbf{i}} f \cdot g \, \mathrm{d}\lambda := \int_{\mathbb{R}^m} f \cdot D^{\mathbf{i}} g \, \mathrm{d}\lambda.$$

Furthermore, we also need the Fourier transform to define the Bessel potential Sobolev spaces.

Definition 3.11 [Fourier transform]
*For $g \in \mathcal{S}$, the **Fourier transform** $\mathcal{F} : \mathcal{S} \to \mathcal{S}$ is defined by*

$$\mathcal{F}(g)(\boldsymbol{\xi}) := \int_{\mathbb{R}^m} g(\mathbf{t}) \exp(-i\mathbf{t}^T\boldsymbol{\xi}) \, \mathrm{d}\mathbf{t}, \tag{3.11}$$

where $i \in \mathbb{C}$ denotes the imaginary unit. For a tempered distribution $f \in \mathcal{S}'$, the Fourier transform $\mathcal{F} : \mathcal{S}' \to \mathcal{S}'$ is defined via $\mathcal{F}(f)(g) := f(\mathcal{F}(g))$.

It can be shown that $\mathcal{F}$ defines an isomorphism on $\mathcal{S}'$ and the inverse Fourier transform $\mathcal{F}^{-1} : \mathcal{S}' \to \mathcal{S}'$ also exists.

Sobolev spaces of isotropic smoothness

The Sobolev norm is given by

$$\|f\|_{H_p^s(\mathbb{R}^m)} := \left\| \mathcal{F}^{-1}\left((1 + \|\boldsymbol{\xi}\|_{\ell_2}^2)^{\frac{s}{2}} \mathcal{F} f\right) \right\|_{L_p(\mathbb{R}^m)}, \tag{3.12}$$

where $\boldsymbol{\xi}$ is used instead of $\mathbf{t}$ as coordinate vector to indicate that we are in the frequency domain. A detailed analysis of the above expression, including the fact that it is well-defined for arbitrary $f \in \mathcal{S}'$, can be found in [7].

Definition 3.12 [Bessel potential Sobolev space]
*Let $1 \leq p \leq \infty$ and $s \geq 0$. The **(Bessel potential) Sobolev space** of isotropic smoothness of order s on $\mathbb{R}^m$ is defined by*

$$H_p^s(\mathbb{R}^m) := \{g \in L_p(\mathbb{R}^m) \mid \|g\|_{H_p^s(\mathbb{R}^m)} < \infty\}. \tag{3.13}$$

With the restriction

$$\|f\|_{H_p^s(T)} := \inf_{g \in H_p^s(\mathbb{R}^m), g|_T = f} \|g\|_{H_p^s(\mathbb{R}^m)} \tag{3.14}$$

for an open bounded domain $T \subset \mathbb{R}^m$, we define the Sobolev space of isotropic smoothness of order s on T by

$$H_p^s(T) := \{g \in L_p(T) \mid \|g\|_{H_p^s(T)} < \infty\}.$$

Alternatively to this definition, the Bessel potential Sobolev spaces can also be defined by complex interpolation of Sobolev-Slobodeckij spaces, see [80] for details. In the special case $s \in \mathbb{N}$ and $1 < p < \infty$, the original Sobolev norm

$$\|f\|_{W_p^s(T)} := \left(\sum_{\mathbf{i} \in \mathbb{N}^m, \|\mathbf{i}\|_{\ell_1} \leq s} \|D^{\mathbf{i}} f\|_{L_p(T)}^p \right)^{\frac{1}{p}} \tag{3.15}$$

is equivalent to $\|\cdot\|_{H_p^s(T)}$. Here, $D^{\mathbf{i}} f \in \mathcal{S}'$ has to be understood in the sense of (3.10). Thus, the norm is finite if and only if $D^{\mathbf{i}} f \in L_p(T) \subset \mathcal{S}'$. For $p = 2$ and any $s \geq 0$, $H_2^s(T)$ is a Hilbert space and we commonly write $H^s(T)$ for $H_2^s(T)$.

Sobolev spaces of mixed smoothness

Besides the Sobolev spaces of isotropic smoothness, also the Sobolev spaces of (dominating) mixed smoothness $H_{\text{mix},p}^s(T)$ will be used in this thesis. On $\mathbb{R}^m$, their norm is defined via

$$\|f\|_{H_{\text{mix},p}^s(\mathbb{R}^m)} := \left\| \mathcal{F}^{-1} \left(\prod_{i=1}^m (1 + |\xi_i|^2)^{\frac{s}{2}} \mathcal{F} f \right) \right\|_{L_2(\mathbb{R}^m)}, \tag{3.16}$$

where ξ_i denotes the i-th coordinate in the frequency domain. By restriction to T, the corresponding norm $\|\cdot\|_{H_{\text{mix},p}^s(T)}$ is defined analogously to (3.14).

Definition 3.13 [Bessel potential Sobolev space of mixed smoothness]
*Let $1 \leq p \leq \infty$ and $s \geq 0$. The **(Bessel potential) Sobolev space of mixed smoothness** of order s on $\mathbb{R}^m$ is defined by*

$$H_{\text{mix},p}^s(T) := \{f \in L_p(T) \mid \|f\|_{H_{\text{mix},p}^s(T)} < \infty\}.$$

With the restriction

$$\|f\|_{H^s_{\mathrm{mix},p}(T)} := \inf_{g \in H^s_{\mathrm{mix},p}(\mathbb{R}^m), g|_T = f} \|g\|_{H^s_{\mathrm{mix},p}(\mathbb{R}^m)}$$

for an open bounded domain $T \subset \mathbb{R}^m$, we define the Sobolev space of mixed smoothness of order s on T by

$$H^s_{\mathrm{mix},p}(T) := \{g \in L_p(T) \mid \|g\|_{H^s_{\mathrm{mix},p}(T)} < \infty\}.$$

Similar to $H^s_p(T)$, we observe that, for the case $s \in \mathbb{N}$ and $1 < p < \infty$, the norm

$$\|f\|_{W^s_{\mathrm{mix},p}(T)} := \left(\sum_{\mathbf{i} \in \mathbb{N}^m, \|\mathbf{i}\|_{\ell_\infty} \leq s} \|D^{\mathbf{i}} f\|^p_{L_p(T)} \right)^{\frac{1}{p}} \tag{3.17}$$

is equivalent to $\|\cdot\|_{H^s_{\mathrm{mix},p}(T)}$, see e.g. [76]. Note that, differently to the norm (3.15), all multiindices $\mathbf{i}$ which fulfill $\|\mathbf{i}\|_{\ell_\infty} \leq s$ instead of $\|\mathbf{i}\|_{\ell_1} \leq s$ are considered. Again, in the case $p = 2$, the space $H^s_{\mathrm{mix}}(T) := H^s_{\mathrm{mix},2}(T)$ is a Hilbert space for arbitrary $s \geq 0$.

3.2.2 Sobolev-Bochner spaces of vector-valued functions

Similar to the vector-valued Lebesgue spaces in subsection 3.1.2, we now consider Sobolev spaces of vector-valued functions. As in the scalar-valued case in subsection 3.2.1, we restrict ourselves to $\omega = \lambda$. We are particularly interested in domains $T \subset \mathbb{R}^m$ with Lipschitz boundary. In principle, the definitions work completely analogously to the scalar-valued case. The only difference is that the $L_p(T; E)$ norms are used instead of the $L_p(T)$ norms, where E denotes a separable Hilbert space. Formally, one has to consider the Schwartz space $\mathcal{S}(\mathbb{R}^m; E)$ of E-valued functions and the space of E-valued tempered distributions. However, we will neglect most technical details here and just present the definitions in terms of bounds on $L_p(T; E)$ norms of vector-valued Gâteaux derivatives. For details on the vector-valued variants of the Schwartz space and the tempered distributions, we refer the reader to [81].

Note that, in the vector-valued case, the norm equivalence between the Bessel potential space and the Slobodeckij space is only valid if E fulfills certain conditions[3], see e.g. [72, 81]. However, these conditions are automatically fulfilled if E is a Hilbert space.

[3] If E has the so-called unconditionality of martingale differences property (UMD-property), i.e. if the Hilbert transform is bounded in $L_2(\mathbb{R}^m; E)$, most of the results from the scalar-valued case carry over to the case of vector-valued functions.

Sobolev-Bochner spaces of isotropic smoothness

Analogously to (3.12), the Sobolev-Bochner norm for a Hilbert space E is given by

$$\|f\|_{H_p^s(\mathbb{R}^m;E)} := \|\mathcal{F}^{-1}\left((1+\|\boldsymbol{\xi}\|_{\ell_2}^2)^{\frac{s}{2}}\mathcal{F}f\right)\|_{L_p(\mathbb{R}^m;E)},$$

where $\mathcal{F}$ now denotes the Fourier transform for functions in $L_2(\mathbb{R}^m;E)$. The formula reads the same[4] as in (3.11). However, the corresponding integral has to be understood as a Bochner integral now. Then, the norm on an open $T \subset \mathbb{R}^m$ is defined by restriction

$$\|f\|_{H_p^s(T;E)} = \inf_{g\in H_p^s(\mathbb{R}^m;E), g|_T=f} \|g\|_{H_p^s(\mathbb{R}^m;E)} \tag{3.18}$$

as in (3.14).

Definition 3.14 [Bessel potential Sobolev-Bochner space]
*Let $1 \le p \le \infty$ and $s \ge 0$. The **(Bessel potential) Sobolev-Bochner space** is defined via*

$$H_p^s(T;E) := \{f \in L_p(T;E) \mid \|f\|_{H_p^s(T;E)} < \infty\}.$$

For $s \in \mathbb{N}$ and $1 < p < \infty$, we again obtain the norm equivalence between the Bessel potential norm of $H_p^s(T;E)$ and the Sobolev-Slobodeckij norm

$$\|f\|_{W_p^s(T;E)} := \left(\sum_{\mathbf{i}\in\mathbb{N}^m, \|\mathbf{i}\|_{\ell_1}\le s} \|D^{\mathbf{i}}f\|_{L_p(T;E)}^p\right)^{\frac{1}{p}},$$

where the weak partial derivatives $D^{\mathbf{i}}$ on elements from the vector-valued Schwartz space $\mathcal{S}(\mathbb{R}^m;E)$ have to be understood in the Gâteaux-sense. For $p=2$, the space $H^s(T;E) := H_2^s(T;E)$ is a Hilbert space for any $s \ge 0$.

Sobolev-Bochner spaces of mixed smoothness

The Sobolev-Bochner spaces of mixed smoothness are also treated similarly as in the scalar-valued case. We define the corresponding Sobolev-Bochner norm for a Hilbert space E by

$$\|f\|_{H^s_{\mathrm{mix},p}(\mathbb{R}^m;E)} := \left\|\mathcal{F}^{-1}\left(\prod_{i=1}^m (1+|\xi_i|^2)^{\frac{s}{2}}\mathcal{F}f\right)\right\|_{L_2(\mathbb{R}^m;E)},$$

where ξ_i represents the i-th coordinate. The norm on an open domain $T \subset \mathbb{R}^m$ with Lipschitz boundary is then defined by the restriction

$$\|f\|_{H^s_{\mathrm{mix},p}(T;E)} = \inf_{g\in H^s_{\mathrm{mix},p}(\mathbb{R}^m;E), g|_T=f} \|g\|_{H^s_{\mathrm{mix},p}(\mathbb{R}^m;E)}.$$

[4] Formally, one has to work with the complexification of the vector space E.

Definition 3.15 [Bessel potential Sobolev-Bochner space of mixed smoothness]
*Let $1 \leq p \leq \infty$ and $s \geq 0$. The **(Bessel potential) Sobolev-Bochner space of mixed smoothness** is defined by*

$$H^s_{\mathrm{mix},p}(T;E) := \{f \in L_p(T;E) \mid \|f\|_{H^s_{\mathrm{mix},p}(T;E)} < \infty\}.$$

As in the scalar-valued case, cf. (3.17), the norm

$$\|f\|_{W^s_{\mathrm{mix},p}(T;E)} := \left(\sum_{\mathbf{i}\in\mathbb{N}^m, \|\mathbf{i}\|_{\ell_\infty}\leq s} \|D^{\mathbf{i}} f\|^p_{L_p(T;E)} \right)^{\frac{1}{p}}$$

is equivalent to $\|\cdot\|_{H^s_{\mathrm{mix},p}(T;E)}$ for $s \in \mathbb{N}$ and $1 < p < \infty$, see [76]. Similar to the case of isotropic smoothness spaces, $H^s_{\mathrm{mix}}(T;E) := H^s_{\mathrm{mix},2}(T;E)$ is a Hilbert space for $p = 2$ and arbitrary $s \geq 0$.

3.3 Vector-valued reproducing kernel Hilbert spaces

As before, let $T \subset \mathbb{R}^m$ be an open Lipschitz domain and let $(E, \langle\cdot,\cdot\rangle_E)$ be a real Hilbert space. Many statements of the later chapters of this thesis are valid for function spaces $\mathcal{H}$ which are continuously embedded into $C(T;E)$, the space of continuous E-valued functions on T equipped with the norm

$$\|f\|_{C(T;E)} := \sup_{\mathbf{t}\in T} \|f(\mathbf{t})\|_E. \tag{3.19}$$

A so-called reproducing kernel Hilbert space (RKHS) $\mathcal{H}$ is one of the most important valid candidates. Therefore, we shortly review the definition of an RKHS and state relevant properties of such a space. To this end, we follow the lines of [58, 63].

3.3.1 RKHS and continuous embeddings into $C(T;E)$

First, we need to consider a kernel function with image in the space $\mathcal{L}(E,E)$ of bounded linear operators from E to E.

Definition 3.16 [Kernel Function]
*A function $K : T \times T \to \mathcal{L}(E,E)$ is called a **kernel function** if the following conditions are fulfilled:*

1. *Symmetry: For each $\mathbf{s},\mathbf{t} \in T$ it holds $K(\mathbf{s},\mathbf{t}) = K(\mathbf{t},\mathbf{s})^*$, where the latter denotes the adjoint operator of $K(\mathbf{t},\mathbf{s})$.*

2. *Non-negativity: For any* $n \in \mathbb{N}$ *and arbitrary points* $\mathbf{t}_i \in T$, $\mathbf{x}_i \in E$, $i = 1, \dots, n$, *it holds*

$$\sum_{i,j=1}^{n} \langle \mathbf{x}_i, K(\mathbf{t}_i, \mathbf{t}_j)(\mathbf{x}_j) \rangle_E \geq 0.$$

In the scalar-valued case $E = \mathbb{R}$, condition 2 implies the positive semidefiniteness of the kernel.

Definition 3.17 [(Vector-valued) Reproducing Kernel Hilbert Space]
Let $(H, \langle \cdot, \cdot \rangle_H)$ *be a Hilbert space of functions from* T *to* E. *We call* H *a* ***reproducing kernel Hilbert space*** *if, for any* $\mathbf{t} \in T$ *and* $\mathbf{x} \in E$, *the point evaluation functional* $I_{\mathbf{t},\mathbf{x}} : H \to \mathbb{R}$ *defined by*

$$I_{\mathbf{t},\mathbf{x}}(f) := \langle \mathbf{x}, f(\mathbf{t}) \rangle_E$$

is continuous.

Note that, due to the Riesz representation theorem, there exists a linear operator $K_{\mathbf{t}} : E \to H$ such that

$$I_{\mathbf{t},\mathbf{x}}(f) = \langle \mathbf{x}, f(\mathbf{t}) \rangle_E = \langle K_{\mathbf{t}}\mathbf{x}, f \rangle_H. \tag{3.20}$$

It is straightforward to see that the function $K : T \times T \to \mathcal{L}(E, E)$ defined by

$$K(\mathbf{s}, \mathbf{t})(\mathbf{x}) := (K_{\mathbf{t}}\mathbf{x})(\mathbf{s}) \in E \tag{3.21}$$

is a kernel function, see e.g. [58] for details. Therefore, each reproducing kernel Hilbert space admits a kernel K defined by the Riesz representer of the point evaluation functional $I_{\mathbf{t},\mathbf{x}}$. Conversely, every kernel K characterizes the corresponding RKHS uniquely by completion of

$$\operatorname{span}\{K(\cdot, \mathbf{t})(\mathbf{x}) \mid \mathbf{t} \in T, \mathbf{x} \in E\} \tag{3.22}$$

with respect to the inner product

$$\langle f, g \rangle_H := \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} \langle \mathbf{x}_i, K(\mathbf{t}_i, \mathbf{s}_j)(\mathbf{y}_j) \rangle_E$$

for arbitrary $n_1, n_2 \in \mathbb{N}$, $f(\cdot) = \sum_{i=1}^{n_1} K(\cdot, \mathbf{t}_i)(\mathbf{x}_i)$ and $g(\cdot) = \sum_{j=1}^{n_2} K(\cdot, \mathbf{s}_j)(\mathbf{y}_j)$, see e.g. [58, 63] for details. This one-to-one relationship between the kernel and the RKHS allows us to prove the following important proposition.

Proposition 3.18
Let $(H, \langle \cdot, \cdot \rangle_H)$ *be a reproducing kernel Hilbert space of functions from* T *to* E *with continuous kernel* $K : T \times T \to \mathcal{L}(E, E)$. *Let there exist a continuous extension of* K *to the closure* $\bar{T} \times \bar{T}$ *of* $T \times T$. *Then the embedding* $(H, \langle \cdot, \cdot \rangle_H) \hookrightarrow (C(T; E), \| \cdot \|_\infty)$ *is continuous.*

Proof. For every $f \in H, \mathbf{t} \in T$ and $\mathbf{x} \in E$ we obtain

$$\begin{aligned}\langle \mathbf{x}, f(\mathbf{t})\rangle_E &\overset{(3.20)}{=} \langle K_{\mathbf{t}}\mathbf{x}, f\rangle_H \leq \|K_{\mathbf{t}}\mathbf{x}\|_H \|f\|_H = \sqrt{\langle K_{\mathbf{t}}\mathbf{x}, K_{\mathbf{t}}\mathbf{x}\rangle_H}\,\|f\|_H \\ &\overset{(3.20)}{=} \sqrt{\langle \mathbf{x}, (K_{\mathbf{t}}\mathbf{x})(\mathbf{t})\rangle_E}\,\|f\|_H \overset{(3.21)}{=} \sqrt{\langle \mathbf{x}, K(\mathbf{t},\mathbf{t})(\mathbf{x})\rangle_E}\,\|f\|_H \\ &\leq \sqrt{\|K(\mathbf{t},\mathbf{t})\|_{\mathcal{L}(E,E)}}\,\|\mathbf{x}\|_E\,\|f\|_H,\end{aligned}$$

where

$$\|L\|_{\mathcal{L}(X,Y)} := \sup_{\|\mathbf{y}\|_X=1} \|L(\mathbf{y})\|_Y \tag{3.23}$$

denotes the standard operator norm for linear operators $L : X \to Y$. Since there exists a continuous extension $\tilde{K} : \bar{T} \times \bar{T} \to \mathcal{L}(E,E)$ of K and since $\|\tilde{K}(\cdot,\cdot)\|_{\mathcal{L}(E,E)}$ is a continuous function on the compact domain $\bar{T} \times \bar{T}$, there exists a $c > 0$ such that

$$\sup_{\mathbf{t}\in T} \sqrt{\|K(\mathbf{t},\mathbf{t})\|_{\mathcal{L}(E,E)}} \leq \max_{\mathbf{t}\in \bar{T}} \sqrt{\|\tilde{K}(\mathbf{t},\mathbf{t})\|_{\mathcal{L}(E,E)}} = c < \infty.$$

Therefore, we obtain

$$\|f(\mathbf{t})\|_E^2 = \langle f(\mathbf{t}), f(\mathbf{t})\rangle_E \leq \sqrt{\|K(\mathbf{t},\mathbf{t})\|_{\mathcal{L}(E,E)}}\,\|f(\mathbf{t})\|_E\,\|f\|_H \leq c\|f(\mathbf{t})\|_E\,\|f\|_H,$$

and thus

$$\|f\|_\infty = \sup_{\mathbf{t}\in T} \|f(\mathbf{t})\|_E \leq c\|f\|_H. \tag{3.24}$$

This shows that convergence with respect to $\|\cdot\|_H$ implies convergence with respect to $\|\cdot\|_\infty$. Since the latter coincides with uniform convergence, a converging sequence of functions from $\operatorname{span}\{K(\cdot,\mathbf{t})\mathbf{x} \mid \mathbf{t} \in T, \mathbf{x} \in E\}$ has a continuous limit function. This is due to the fact that $K(\cdot,\mathbf{t})\mathbf{x}$ is a continuous function for every $\mathbf{t} \in T$ and $\mathbf{x} \in E$ because of the continuity of K. Therefore, all functions in H are elements of $C(T;E)$ and the embedding $(H, \langle\cdot,\cdot\rangle_H) \hookrightarrow (C(T;E), \|\cdot\|_\infty)$ is continuous by (3.24). □

3.3.2 Examples

We conclude our review on vector-valued reproducing kernel Hilbert spaces with two examples.

Componentwise RKHS

Let H be an RKHS of functions from T to $E = \mathbb{R}^d$. In this case, the space $\mathcal{L}(E,E)$ consists of all quadratic matrices in $\mathbb{R}^{d\times d}$. Furthermore, if each component function of the RKHS H additionally belongs to a scalar-valued RKHS, the standard theory for scalar-valued reproducing kernel Hilbert spaces directly carries over to the vector-valued case. To this end, let $K_1, \ldots, K_d : T \times T \to \mathbb{R}$ be d scalar-valued kernels. The corresponding

matrix-valued kernel is given by

$$K(\mathbf{s},\mathbf{t}) = \operatorname{diag}\left(K_1(\mathbf{s},\mathbf{t}),\ldots,K_d(\mathbf{s},\mathbf{t})\right). \tag{3.25}$$

Indeed, a straightforward evaluation of (3.22) shows that the space corresponding to K is given by the completion of

$$\begin{aligned} &\operatorname{span}\left\{\left.\begin{pmatrix} K_1(\cdot,\mathbf{t})x_1 \\ \vdots \\ K_d(\cdot,\mathbf{t})x_d \end{pmatrix}\right| \mathbf{t}\in T, \mathbf{x}\in\mathbb{R}^d\right\} = \operatorname{span}\left\{K_1(\cdot,\mathbf{t})\mathbf{e}_1,\ \ldots,\ K_d(\cdot,\mathbf{t})\mathbf{e}_d | \mathbf{t}\in T\right\} \\ = \ &\operatorname{span}\left\{K_1(\cdot,\mathbf{t})\mathbf{e}_1 \mid \mathbf{t}\in T\right\} \oplus \ldots \oplus \operatorname{span}\left\{K_d(\cdot,\mathbf{t})\mathbf{e}_d \mid \mathbf{t}\in T\right\}. \end{aligned}$$

Sobolev spaces

Let $m = 1, T = (0,1)$ and $\mathbb{E} = \mathbb{R}$. The Sobolev space $H^1(T)$ is a reproducing kernel Hilbert space. However, the closed form expression of the kernel function of $H^1(T)$ is quite involved. If we restrict ourselves to the space $H^1_{2,0}(T) = \{f \in H^1(T) \cap C(\bar{T}) \mid f(0) = f(1) = 0\}$, the kernel corresponding to the norm (3.15) is given by

$$K(s,t) = \frac{\sinh(\min(s,t))\sinh(1-\max(s,t))}{\sinh(1)}.$$

For the derivation of this expression and other kernel formulas for certain Sobolev spaces, we refer to [18]. Details on the spaces $H^1_{2,0}(T)$, which can be defined via the trace operator in the higher-dimensional case, can be found in [1].

In the multivariate case, the Sobolev spaces $H^s(T)$ with $T = (0,1)^m, s > 0, m \geq 1$ are reproducing kernel Hilbert spaces only if $s > \frac{m}{2}$. This is due to the Sobolev embedding theorem, see e.g. theorem 4.6.1 of [80]. Since the Sobolev spaces of dominating mixed smoothness $H^s_{\text{mix}}(T)$ can be defined as a tensor product of m univariate spaces $H^s((0,1))$, see [52] for details, it can be shown in a straightforward manner that $H^s_{\text{mix}}(T)$ is a reproducing kernel space for all $s > \frac{1}{2}$, independent of the dimension m. The corresponding kernel is the product of m kernels of the univariate space $H^s((0,1))$. It is noteworthy that the extension of this kernel to $\bar{T} = [0,1]^m$ is continuous and therefore proposition 3.18 can be applied.

The vector-valued Sobolev-Bochner spaces $H^s(T;\mathbb{R}^d)$ for $s > \frac{m}{2}$ and $H^s_{\text{mix}}(T;\mathbb{R}^d)$ for $s > \frac{1}{2}$ are then constructed as componentwise reproducing kernel Hilbert spaces as explained above.

3.4 Interpolation spaces

When we discuss vector-valued regression in infinite-dimensional function spaces in chapter 4, interpolation theory will play an important role, see e.g. [7, 24, 67] for an overview.

Therefore, we provide a short introduction into real interpolation theory based on the so-called $\mathbb{K}$-functional.

3.4.1 Interpolation spaces and the $\mathbb{K}$-functional

The following definition is according to [7, 23].

Definition 3.19 [Real Interpolation Space and $\mathbb{K}$-functional]
*Let $(X, \|\cdot\|_X)$ and $(Y, \|\cdot\|_Y)$ be Banach spaces and let $Y \subset X$. We define the **(real) interpolation space** $(X,Y)_\sigma$ for $\sigma \in (0,1)$ by*

$$(X,Y)_\sigma := \{f \in X \mid \|f\|_\sigma < \infty\}$$

with

$$\|f\|_\sigma := \sup_{t>0} \frac{\mathbb{K}(f,t)}{t^\sigma},$$

*where the so-called $\mathbb{K}$-**functional** $\mathbb{K} : X \times \mathbb{R} \to \mathbb{R}$ is given by*

$$\mathbb{K}(f,t) := \inf_{g \in Y} (\|f-g\|_X + t\|g\|_Y). \tag{3.26}$$

Note that $\mathbb{K}(f,t)$ is essentially the best approximation error to f by a function from Y for small $t > 0$. If $\|f\|_\sigma < \infty$, the decay rate of this best approximation error is at least t^σ.

The following theorem describes how fast the error of the best approximation by a norm-bounded function decays. It will prove to be very helpful when dealing with constrained regression.

Theorem 3.20 [Best Approximation in Norm Balls]
Let $(X, \|\cdot\|_X)$ and $(Y, \|\cdot\|_Y)$ be Banach spaces with $Y \subset X$. Let, furthermore, the embedding $Y \hookrightarrow X$ be continuous. If $f \in (X,Y)_\sigma$ for a $\sigma \in (0,1)$, then

$$\inf_{\|g\|_Y \leq b} \|f-g\|_X \leq b^{-\frac{\sigma}{1-\sigma}} \cdot \|f\|_\sigma^{\frac{1}{1-\sigma}}$$

for all $b > 0$.

Proof. See theorem 4.16 in [23]. □

A slightly more general statement and the corresponding proof can also be found in [23].

3.4.2 Examples

To illustrate the concept of real interpolation spaces, we now provide an example in terms of interpolation scales of Sobolev spaces. The resulting function spaces are certain

Besov spaces. To define them, let $\Phi \in \mathcal{S}(\mathbb{R}^m)$ be a function from the Schwartz space such that $\Phi(\mathbf{t}) = 1$ for all $\|\mathbf{t}\|_{\ell_2} \leq 1$ and $\Phi(\mathbf{t}) = 0$ for all $\|\mathbf{t}\|_{\ell_2} \geq \frac{3}{2}$. Let furthermore

$$\Phi_0 = \Phi, \quad \Phi_1(\mathbf{t}) = \Phi\left(\frac{\mathbf{t}}{2}\right) - \Phi(\mathbf{t}) \quad \text{and} \quad \Phi_k(\mathbf{t}) = \Phi_1\left(2^{1-k}\mathbf{t}\right) \ \forall\, k \geq 2$$

hold for all $\mathbf{t} \in \mathbb{R}^m$. It is easy to see that these functions fulfill

$$\sum_{k=0}^{\infty} \Phi_k(\mathbf{t}) = 1 \quad \forall\, \mathbf{t} \in \mathbb{R}^m.$$

Therefore, they are called *dyadic resolution of unity*.

Definition 3.21 [Besov spaces of isotropic smoothness]
Let E be a separable Hilbert space and let $(\Phi_k)_{k=0}^{\infty}$ be the dyadic resolution of unity from above. Let furthermore $s > 0$ and $1 \leq p < \infty$ and $1 \leq q \leq \infty$. The ***Besov space*** *$B^s_{p,q}(\mathbb{R}^m; E)$ is defined by*

$$B^s_{p,q}(\mathbb{R}^m; E) := \left\{ f \in L_p(\mathbb{R}^m; E) \mid \|f\|_{B^s_{p,q}(\mathbb{R}^m;E)} < \infty \right\},$$

where

$$\|f\|_{B^s_{p,q}(\mathbb{R}^m;E)} := \left\| \left(2^{sj} \|\mathcal{F}^{-1}\left(\Phi_j \cdot \mathcal{F} f\right)\|_{L_p(\mathbb{R}^m;E)} \right)_{j=0}^{\infty} \right\|_{\ell_q}.$$

The Besov spaces on a domain $T \subset \mathbb{R}^m$ are defined by

$$B^s_{p,q}(T; E) := \left\{ f \in L_p(\mathbb{R}^m; E) \mid \exists\, g \in B^s_{p,q}(\mathbb{R}^m; E) \text{ such that } f = g|_T \right\}$$

with norm

$$\|f\|_{B^s_{p,q}(T;E)} := \inf_{g \in B^s_{p,q}(\mathbb{R}^m;E), f = g|_T} \|g\|_{B^s_{p,q}(\mathbb{R}^m;E)}.$$

For details on Besov spaces and vector-valued distributions, see e.g. [7, 64, 72, 80, 82]. In the special case $p = q = 2$, one can show that $B^s_{2,2}(T; E) = H^s_2(T; E)$, see e.g. [80]. We can now state the result on the interpolation between Sobolev-Bochner spaces.

Theorem 3.22 [Interpolation between Sobolev-Bochner spaces]
Let $0 < \sigma < 1$, let $s_2 > s_1 \geq 0$ and let $1 < p < \infty$. For an open Lipschitz domain $T \subset \mathbb{R}^m$ and a separable Hilbert space E, we obtain

$$\left(H^{s_1}_p(T; E), H^{s_2}_p(T; E) \right)_{\sigma} = B^{(1-\sigma)s_1 + \sigma s_2}_{p,\infty}(T; E),$$

where the equality has to be understood in the sense of equal sets and norm equivalence of the corresponding spaces.

Proof. See e.g. section 4.3.1 of [80] for the scalar-valued case. For the vector-valued case, combine theorems 6 and 8 of [64] with the norm equivalences for Besov spaces in [82]. □

Although the resulting spaces in theorem 3.22 are no longer Sobolev-Bochner spaces, we have $H_2^s(T;E) \subset B_{2,\infty}^s(T;E)$ for all $s > 0$ in the special case $p = 2$. This shows that

$$\left(H_2^{s_1}(T;E), H_2^{s_2}(T;E)\right)_\sigma \supset H_2^{(1-\sigma)s_1+\sigma s_2}(T;E). \tag{3.27}$$

We refer to [64] and section 4.6 of [80] for more details.

It is noteworthy that there exist canonical definitions for Besov spaces of dominating mixed smoothness, see [82]. Furthermore, there are similar interpolation results for these spaces if the domain is $\mathbb{R}^m$, see theorem 3.9 of [73]. However, the question whether or not these results are also valid on bounded Lipschitz domains has only been answered for very specific choices of parameters up to now. The main reason for this is that one has to ensure the existence of a so-called *common extension operator*. We refer the interested reader to section 1.17 of [80] and to sections 1.2.8 and 3.2.4 of [82], where several special cases are considered and a conjecture for the Sobolev-Bochner spaces can be found.

3.5 Full grids, sparse grids and hyperbolic crosses

For many applications, such as solving partial differential equations, uncertainty quantification and machine learning, grid-based discretizations are commonly used. In data mining for instance, they have proven to be a good alternative to data-based approaches such as support vector machines, see [74], where the number of degrees of freedom - and often also the shape and location of the corresponding basis functions - is coupled to the input data. Most notably, if the number of data is large, as in most Big Data applications, see e.g. [11, 15, 35], grid-based methods prevail. We will comment on the drawbacks of data-based approaches in this situation in more detail in section 5.1.2.

If the dimension of the space from which the data stems is larger than 3, conventional *full grid* tensor-product methods are no longer feasible since they suffer from the *curse of dimensionality*, see [5], i.e. the exponential growth of the degrees of freedom with increasing dimension. To overcome this curse, at least to some extent, we can employ *hyperbolic crosses*, see [78], and *sparse grids*, see [14].

Based on [14] and [78], we briefly review the concepts of sparse grids and hyperbolic crosses, which will serve as examples to illustrate our results in the later chapters. For reasons of clarity and comprehensibility, we restrict ourselves to linear splines on sparse grids and Fourier polynomials on hyperbolic crosses. Note, however, that these constructions work analogously for different bases such as higher order splines or global polynomials for instance, see also [14, 27].

3.5.1 Linear prewavelet splines on full grids and sparse grids

We now introduce the so-called piecewise linear *prewavelet* basis, see [41]. In contrast to the hat function basis, which is commonly used in sparse grid applications because of

its simplicity, see e.g. [11, 15, 31], the main advantage of the prewavelet basis is that it forms a Riesz frame with respect to the $L_2(T)$ norm, which we will exploit several times.

The prewavelet basis

We define the univariate hat function $\phi : \mathbb{R} \to \mathbb{R}$ by

$$\phi(t) := \begin{cases} 1 - |t| & \text{if } t \in [-1, 1], \\ 0 & \text{else} \end{cases}$$

and its translations and dilations $\phi_{l,i} : [0, 1] \to \mathbb{R}$ by

$$\phi_{l,i}(t) := \phi(2^l \cdot t - i)|_{[0,1]} \tag{3.28}$$

for $l \in \mathbb{N}$ and $i \in \{0, 1, \ldots, 2^l - 1, 2^l\}$. The univariate prewavelets $\gamma_{l,i} : [0, 1] \to \mathbb{R}$ can now be constructed as in [41]:

$$\gamma_{0,0} := 1, \ \gamma_{0,1} := \phi_{0,1}, \ \gamma_{1,1} := 2 \cdot \phi_{1,1} - 1.$$

For $l \geq 2$, let $I_l := \left\{ i \in \mathbb{N} \mid 1 \leq i \leq 2^l - 1, \ i \text{ odd} \right\}$ and let

$$\gamma_{l,i} := 2^{\frac{l}{2}} \cdot \left(\frac{1}{10}\phi_{l,i-2} - \frac{6}{10}\phi_{l,i-1} + \phi_{l,i} - \frac{6}{10}\phi_{l,i+1} + \frac{1}{10}\phi_{l,i+2} \right)$$

for $i \in I_l, i \notin \{1, 2^l - 1\}$ and

$$\gamma_{l,1} := 2^{\frac{l}{2}} \cdot \left(-\frac{6}{5}\phi_{l,0} + \frac{11}{10}\phi_{l,1} - \frac{3}{5}\phi_{l,2} + \frac{1}{10}\phi_{l,3} \right), \quad \gamma_{l,2^l-1}(t) := \gamma_{l,1}(1 - t).$$

A depiction can be found in figure 3.1. The m-variate prewavelet functions are then defined by a tensor product approach

$$\gamma_{\mathbf{l},\mathbf{i}}(\mathbf{t}) := \prod_{j=1}^{m} \gamma_{l_j,i_j}(t_j) \tag{3.29}$$

with the multivariate level index $\mathbf{l} = (l_1, \ldots, l_m) \in \mathbb{N}^m$ and the multivariate position index $\mathbf{i} = (i_1, \ldots, i_m) \in \mathbb{N}^m$.

Full grids and sparse grids

We define

$$\mathbf{I}_\mathbf{l} := \left\{ \mathbf{i} \in \mathbb{N}^m \;\middle|\; \begin{array}{ll} 0 \leq i_j \leq 1, & \text{if } l_j = 0, \\ 1 \leq i_j \leq 2^{l_j} - 1, \ i_j \text{ odd} & \text{if } l_j > 0 \end{array} \quad \text{for all } 1 \leq j \leq m \right\}. \tag{3.30}$$

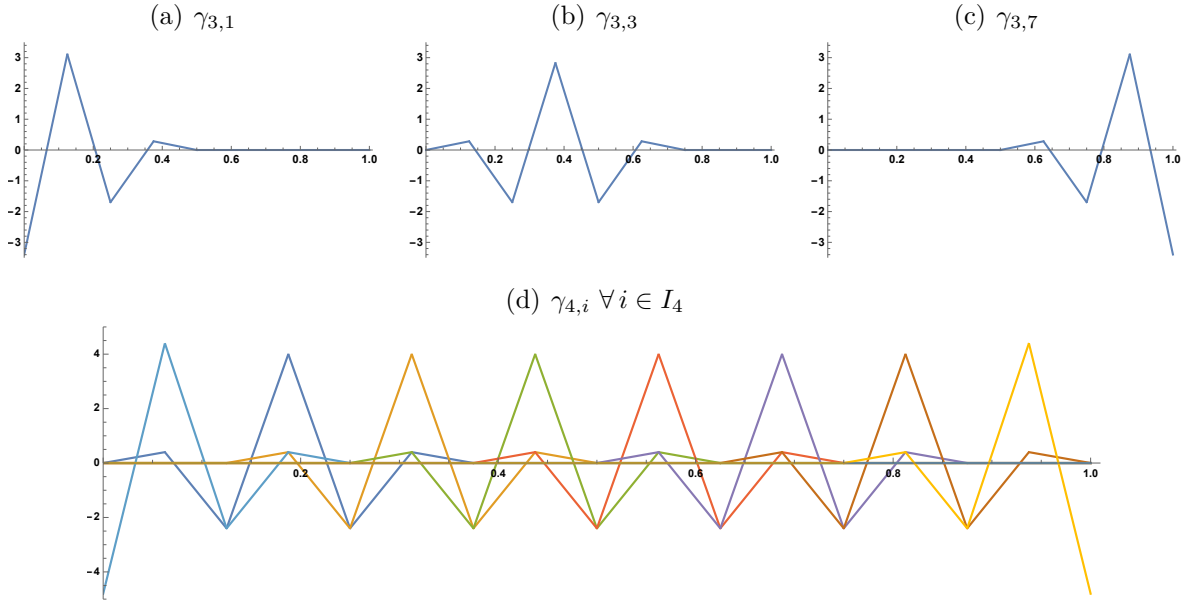

Fig. 3.1: Different univariate piecewise linear prewavelets.

With this, we are able to write the so-called hierarchical increment space (or detail space) of level $\mathbf{l}$ as

$$W_{\mathbf{l}} := \operatorname{span}\left\{\gamma_{\mathbf{l},\mathbf{i}} \mid \mathbf{i} \in \mathbf{I}_{\mathbf{l}}\right\}. \tag{3.31}$$

In the particular case of the prewavelet basis, the spaces $W_{\mathbf{l}}$ are orthogonal to each other in $L_2((0,1)^m)$, i.e.

$$W_{\mathbf{l}} \perp_{L_2} W_{\mathbf{k}} \quad \forall\, \mathbf{l} \neq \mathbf{k}.$$

The multivariate prewavelet space of functions up to level $\mathbf{l}$ is defined by

$$V_{\mathbf{l}} := \bigoplus_{\mathbf{k} \leq \mathbf{l}} W_{\mathbf{k}} = \operatorname{span}\left\{B_{\mathbf{l}}\right\} \tag{3.32}$$

with the *hierarchical* basis

$$B_{\mathbf{l}} := \left\{\gamma_{\mathbf{k},\mathbf{i}} \mid \mathbf{i} \in \mathbf{I}_{\mathbf{k}}, \mathbf{k} \leq \mathbf{l}\right\}.$$

Here, $\mathbf{k} \leq \mathbf{l}$ is meant elementwise. The full grid space of level $k > 0$ can now be written as

$$\mathcal{V}_k^{\text{full}} := \bigoplus_{\substack{\mathbf{l} \in \mathbb{N}^m \\ |\mathbf{l}|_{\ell_\infty} \leq k}} W_{\mathbf{l}}. \tag{3.33}$$

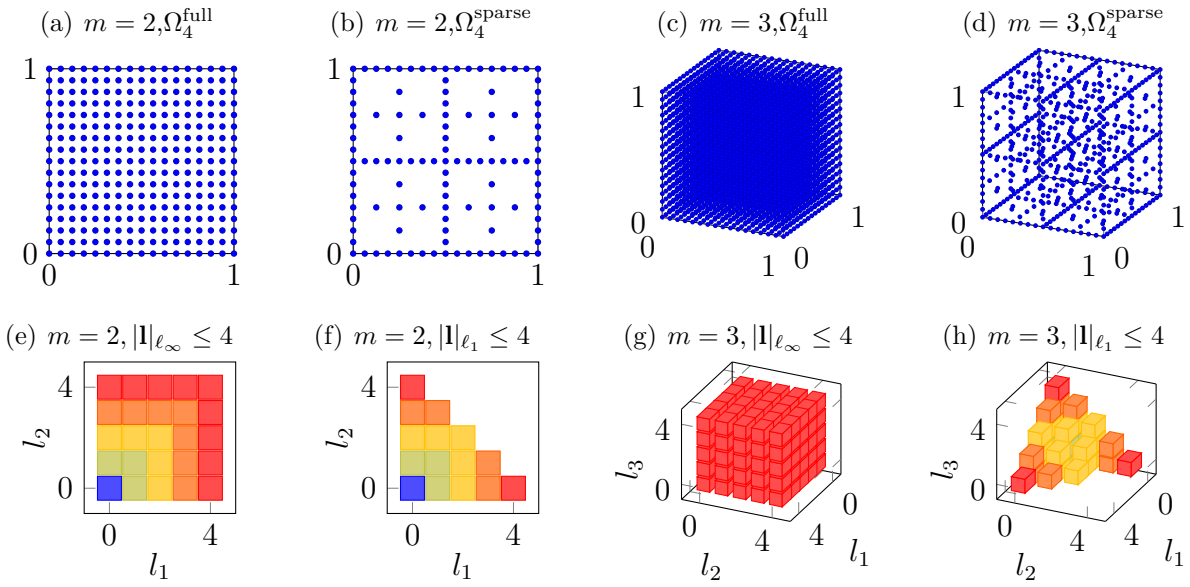

Fig. 3.2: Full and sparse grids of level $k = 4$ in dimensions $m = 2, 3$ (top) and their corresponding index sets (bottom).

Finally, we define the corresponding sparse grid space of level $k > 0$ by

$$\mathcal{V}_k^{\text{sparse}} := \bigoplus_{\substack{\mathbf{l} \in \mathbb{N}^m \\ \zeta_m(\mathbf{l}) \leq k}} W_{\mathbf{l}} \tag{3.34}$$

with $\zeta_m(\mathbf{0}) := 0$ and

$$\zeta_m(\mathbf{l}) := |\mathbf{l}|_{\ell_1} - m + |\{j \mid l_j = 0\}| + 1$$

for every non-zero $\mathbf{l} \in \mathbb{N}^m$. Our specific choice of $\zeta_m(\mathbf{l})$ ensures that the maximum level of sub-grids on the boundary is the same as the maximum level of sub-grids in the interior of the domain. The grids Ω_k^{full} and Ω_k^{sparse}, i.e. the centers of the support of the basis functions belonging to the full grid space and the sparse grid space, can be found in figure 3.2. The corresponding pictures also illustrate the curse of dimensionality for full grids.

Note, that our definitions above only cover the case of scalar-valued functions so far. In the $\mathbb{R}^d$-valued case, we just construct the grid spaces componentwise, i.e. the vector-valued grid space of level $k \in \mathbb{N}$ is defined by

$$\mathcal{V}_k^{*,d} := \text{span}\left\{\gamma_{\mathbf{l},\mathbf{i}} \mathbf{e}_j \mid \gamma_{\mathbf{l},\mathbf{i}} \in \mathcal{V}_k^*,\ j = 1, \ldots, d\right\}, \tag{3.35}$$

where $*$ stands for either "full" or "sparse". Analogously, one can define vector-valued grids where the image space E is an arbitrary separable Banach space by replacing $\mathbf{e}_j$ in the above definition by a basis of that space.

As we mentioned earlier, full grid spaces suffer from the curse of dimensionality. This is reflected in the degrees of freedom

$$\dim\left(\mathcal{V}_k^{\text{full},d}\right) = d \cdot (2^k + 1)^m = d \cdot \mathcal{O}\left(2^{km}\right),$$

which depend exponentially on the dimension m of the domain. The $\mathcal{O}$-notation has to be understood for $k \to \infty$ here, i.e. there exists a (possibly m-dependent) constant $c > 0$ such that $\dim\left(\mathcal{V}_k^{\text{full},d}\right) \leq dc2^{km}$. Conversely, we have

$$\dim\left(\mathcal{V}_k^{\text{sparse},d}\right) = d \cdot \mathcal{O}\left(2^k k^{m-1}\right) \tag{3.36}$$

for $\mathcal{V}_k^{\text{sparse},d}$, see e.g. [30]. Thus, the curse of dimensionality only appears with respect to k instead of 2^k. Therefore, sparse grids are a suitable discretization also for $m > 3$.

Norm equivalences and inverse inequalities

Let $T = (0,1)^m$. An important property which is inherent to the multilevel decomposition of the prewavelet basis is the following norm equivalence for Sobolev spaces:

$$\|f\|^2_{H^s(T;\mathbb{R}^d)} \simeq \sum_{\mathbf{l}\in\mathbb{N}^m} 2^{2s|\mathbf{l}|_{\ell_\infty}} \|w_{\mathbf{l}}\|^2_{L_2(T;\mathbb{R}^d)} \simeq \sum_{\mathbf{l}\in\mathbb{N}^m} 2^{2s|\mathbf{l}|_{\ell_\infty}} \sum_{\mathbf{i}\in\mathbf{I}_{\mathbf{l}}} \sum_{j=1}^{d} |\alpha_{\mathbf{l},\mathbf{i},j}|^2 \quad \forall\, 0 \leq s < \frac{3}{2} \tag{3.37}$$

holds for all $f \in H^s(T;\mathbb{R}^d)$, where $f = \sum_{\mathbf{l}\in\mathbb{N}^m} w_{\mathbf{l}}$ is the unique decomposition of f with respect to the $L_2(T;\mathbb{R}^d)$-orthogonal subspace system given by $\{\bigoplus_{j=1}^d W_{\mathbf{l}}\mathbf{e}_j\}_{\mathbf{l}\in\mathbb{N}^m}$ and we have

$$w_{\mathbf{l}} = \sum_{\mathbf{i}\in\mathbf{I}_{\mathbf{l}}} \left(\alpha_{\mathbf{l},\mathbf{i},1}\, \gamma_{\mathbf{l},\mathbf{i}}, \ldots, \alpha_{\mathbf{l},\mathbf{i},d}\, \gamma_{\mathbf{l},\mathbf{i}}\right)^T.$$

For the case of mixed Sobolev-Bochner spaces, we have

$$\|f\|^2_{H^s_{\text{mix}}(T;\mathbb{R}^d)} \simeq \sum_{\mathbf{l}\in\mathbb{N}^m} 2^{2s|\mathbf{l}|_{\ell_1}} \|w_{\mathbf{l}}\|^2_{L_2(T;\mathbb{R}^d)} \simeq \sum_{\mathbf{l}\in\mathbb{N}^m} 2^{2s|\mathbf{l}|_{\ell_1}} \sum_{\mathbf{i}\in\mathbf{I}_{\mathbf{l}}} \sum_{j=1}^{d} |\alpha_{\mathbf{l},\mathbf{i},j}|^2 \quad \forall\, 0 \leq s < \frac{3}{2}. \tag{3.38}$$

The $\mathbb{R}^d$-valued case we treated here is a trivial consequence of the scalar-valued case of these results, which can be found in [41] and [42]. Note that we directly obtain that the prewavelet basis is a Riesz basis of $L_2(T;\mathbb{R}^d)$ by taking $s = 0$ in the above equations.

Another direct consequence of the norm equivalences above are the corresponding inverse, Bernstein-type inequalities

$$\|f\|^2_{H^s(T;\mathbb{R}^d)} \overset{(3.37)}{\simeq} \sum_{|\mathbf{l}|_{\ell_\infty}\leq k} 2^{2s|\mathbf{l}|_{\ell_\infty}} \|w_{\mathbf{l}}\|^2_{L_2(T;\mathbb{R}^d)} \lesssim 2^{2sk} \sum_{|\mathbf{l}|_{\ell_\infty}\leq k} \|w_{\mathbf{l}}\|^2_{L_2(T;\mathbb{R}^d)} \overset{(3.37)}{\simeq} 2^{2sk} \|f\|^2_{L_2(T;\mathbb{R}^d)} \tag{3.39}$$

for all $f \in \mathcal{V}_k^{\text{full},d}$ and

$$\|f\|^2_{H^s_{\text{mix}}(T;\mathbb{R}^d)} \overset{(3.38)}{\simeq} \sum_{\zeta_m(\mathbf{l})\leq k} 2^{2s|\mathbf{l}|_{\ell_1}} \|w_{\mathbf{l}}\|^2_{L_2(T;\mathbb{R}^d)} \lesssim 2^{2sk} \sum_{\zeta_m(\mathbf{l})\leq k} \|w_{\mathbf{l}}\|^2_{L_2(T;\mathbb{R}^d)} \overset{(3.38)}{\simeq} 2^{2sk} \|f\|^2_{L_2(T;\mathbb{R}^d)} \tag{3.40}$$

for all $f \in \mathcal{V}_k^{\text{sparse},d}$, where $0 \leq s < \frac{3}{2}$. Here, $\lesssim$ means $\leq$ up to a constant which does not depend on k.

Best approximation error

As we will see in the later chapters, the best approximation error measured in the $L_2(T;\mathbb{R}^d)$ norm will be of special interest to derive estimates for the regression error. In the full grid case, we have

$$\inf_{f\in\mathcal{V}_k^{\text{full},d}} \|f-g\|^2_{L_2(T;\mathbb{R}^d)} \lesssim d \cdot 2^{-2sk} \|g\|^2_{H^s(T;\mathbb{R}^d)} \quad \text{if } g \in H^s(T;\mathbb{R}^d) \text{ with } 0 < s \leq 2, \tag{3.41}$$

where the constant involved in the $\lesssim$ estimate only depends on s and m. The prefactor d shows up because one needs to employ the upper bound for the scalar-valued case d times to obtain the vector-valued estimate. In the case of sparse grids, we only obtain

$$\inf_{f\in\mathcal{V}_k^{\text{sparse},d}} \|f-g\|^2_{L_2(T;\mathbb{R}^d)} \lesssim d \cdot 2^{-\frac{2sk}{m}} \|g\|^2_{H^s(T;\mathbb{R}^d)} \quad \text{if } g \in H^s(T;\mathbb{R}^d) \text{ with } 0 < s \leq 2,$$

i.e. although the curse of dimensionality only has a mild effect on the size of the sparse grid, the rate of convergence of the best approximation error deteriorates exponentially with increasing dimension m. For more details on these results, we refer to [42, 52]. However, if additional smoothness is present, i.e. in the case of mixed Sobolev regularity, the exponential dependence on m in the upper bound on the best approximation error is again only present with respect to the level k. To prove this, we first need two combinatorial lemmata.

Lemma 3.23
Let $i, m \in \mathbb{N}$ such that $i > m$. The size of the set $\{\mathbf{l} \in \mathbb{N}^m \mid |\mathbf{l}|_{\ell_1} + |\{j \mid l_j = 0\}| = i\}$ is

$$\sum_{n=0}^{m-1} \binom{i-1-n}{m-1-n} \binom{m}{n}.$$

Proof. Let $\mathbb{N}_+ = \mathbb{N} \setminus \{0\}$. Note that the size of the set $\left\{\mathbf{l} \in \mathbb{N}_+^k \mid |\mathbf{l}|_{\ell_1} = l\right\}$ is $\binom{l-1}{k-1}$ for all choices of $k, l \in \mathbb{N}_+$. Therefore, we have

$$|\{\mathbf{l} \in \mathbb{N}^m \mid |\mathbf{l}|_{\ell_1} + |\{j \mid l_j = 0\}| = i\}| = \sum_{n=0}^{m-1} |\{\mathbf{l} \in \mathbb{N}^m \mid |\mathbf{l}|_{\ell_1} = i-n \wedge |\{j \mid l_j = 0\}| = n\}|$$

$$= \sum_{n=0}^{m-1} \left| \left\{ \mathbf{l} \in \mathbb{N}_+^{m-n} \mid |\mathbf{l}|_{\ell_1} = i - n \right\} \right| \cdot \binom{m}{n}$$
$$= \sum_{n=0}^{m-1} \binom{i-n-1}{m-n-1} \binom{m}{n},$$

which concludes the proof. □

Lemma 3.24
Let $0 < x < 1$ and let $k, m \in \mathbb{N}_+$. It holds

$$\sum_{i=0}^{\infty} x^i \binom{k+i+m-1}{m-1} = \sum_{i=0}^{m-1} \binom{k+m-1}{i} \left(\frac{x}{1-x}\right)^{m-1-i} \cdot \frac{1}{1-x}.$$

Proof. The derivation of this result can be found in the proof of lemma 3.7 of [14]. □

With these auxiliary results, we are able to prove the following theorem on the best approximation error.

Theorem 3.25 [Best approximation error of sparse grids]
For $k \geq 1$, it holds

$$\inf_{f \in \mathcal{V}_k^{\text{sparse},d}} \|f - g\|_{L_2(T;\mathbb{R}^d)}^2 \lesssim d \cdot 2^{-2sk} k^{m-1} \|g\|_{H^s_{\text{mix}}(T;\mathbb{R}^d)}^2 \quad \text{if } g \in H^s_{\text{mix}}(T;\mathbb{R}^d) \text{ with } 0 < s \leq 2. \tag{3.42}$$

Proof. Since we deal with componentwise approximation, the vector-valued result follows directly from the special case $d = 1$. To this end, note that the slightly larger bound with $2^{-2sk} k^{2(m-1)}$ for the rate has already been proven in [42]. To prove (3.42), we essentially follow the lines of the proof of Proposition 6 of [42] but improve on the estimate in the very last step. To this end, let $0 < s \leq 2$ and let $g \in H^s_{\text{mix}}(T)$. Note that

$$\inf_{f \in \mathcal{V}_k^{\text{sparse}}} \|f - g\|_{L_2(T)}^2 \overset{(3.38)}{\simeq} \inf_{f \in \mathcal{V}_k^{\text{sparse}}} \sum_{\zeta_m(\mathbf{l}) \leq k} \|w_{\mathbf{l}}^{(f)} - w_{\mathbf{l}}^{(g)}\|_{L_2(T;\mathbb{R}^d)}^2 + \sum_{\zeta_m(\mathbf{l}) > k} \|w_{\mathbf{l}}^{(g)}\|_{L_2(T;\mathbb{R}^d)}^2,$$

where $f = \sum_{\zeta_m(\mathbf{l}) \leq k} w_{\mathbf{l}}^{(f)}$ and $g = \sum_{\mathbf{l} \in \mathbb{N}^m} w_{\mathbf{l}}^{(g)}$. Thus, by choosing $w_{\mathbf{l}}^{(f)} = w_{\mathbf{l}}^{(g)}$ for all $\mathbf{l}$ with $\zeta_m(\mathbf{l}) \leq k$ and applying

$$\|w_{\mathbf{l}}^{(g)}\|_{L_2(T)} \lesssim 2^{-s|\mathbf{l}|_{\ell_1}} \|g\|_{H^s_{\text{mix}}(T)} \quad \forall \mathbf{l} \in \mathbb{N}^m,$$

which has been proven in [42] for example, we obtain

$$\inf_{f \in \mathcal{V}_k^{\text{sparse}}} \|f - g\|_{L_2(T)}^2 \lesssim \|g\|_{H^s_{\text{mix}}(T)}^2 \sum_{\zeta_m(\mathbf{l}) > k} 2^{-2s|\mathbf{l}|_{\ell_1}}.$$

Estimating the sum on the right hand side gives

$$
\begin{aligned}
\sum_{\zeta_m(\mathbf{l})>k} 2^{-2s|\mathbf{l}|_{\ell_1}} &= \sum_{|\mathbf{l}|_{\ell_1}+|\{j|l_j=0\}|>k+m-1} 2^{-2s|\mathbf{l}|_{\ell_1}} \\
&= \sum_{|\mathbf{l}|_{\ell_1}+|\{j|l_j=0\}|>k+m-1} 2^{-2s\left(|\mathbf{l}|_{\ell_1}+|\{j|l_j=0\}|\right)} \cdot \underbrace{2^{2s|\{j|l_j=0\}|}}_{\leq 2^{2sm}} \\
&\lesssim \sum_{i=k+m}^{\infty} 2^{-2si} \sum_{|\mathbf{l}|_{\ell_1}+|\{j|l_j=0\}|=i} 1 \\
&\overset{\text{Lemma 3.23}}{=} \sum_{i=k+m}^{\infty} 2^{-2si} \sum_{n=0}^{m-1} \binom{i-1-n}{m-1-n}\binom{m}{n} \\
&= 2^{-2s(k+m)} \sum_{i=0}^{\infty} 2^{-2si} \sum_{n=0}^{m-1} \binom{k+m+i-1-n}{m-1-n}\binom{m}{n} \\
&\lesssim 2^{-2sk} \sum_{i=0}^{\infty} 2^{-2si} \binom{k+m+i-1}{m-1} \underbrace{\sum_{n=0}^{m-1} \binom{m}{n}}_{<2^m} \\
&\lesssim 2^{-2sk} \sum_{i=0}^{\infty} 2^{-2si} \binom{k+m+i-1}{m-1} \\
&\overset{\text{Lemma 3.24}}{=} 2^{-2sk} \sum_{i=0}^{m-1} \binom{k+m-1}{i} \left(\frac{2^{-2s}}{1-2^{-2s}}\right)^{m-1-i} \cdot \frac{1}{1-2^{-2s}} \\
&\lesssim 2^{-2sk} \sum_{i=0}^{m-1} \binom{k+m-1}{i} \\
&= 2^{-2sk} \sum_{i=0}^{m-1} \prod_{j=1}^{i} \frac{k+m-j}{j} \\
&\lesssim 2^{-2sk} \sum_{i=0}^{m-1} \prod_{j=1}^{i} k+m-1 \\
&\lesssim 2^{-2sk} k^{m-1}.
\end{aligned}
$$

The $\lesssim$ constant only depends on s and m. Now, (3.42) follows by using the scalar-valued estimate for every component function. □

Note that similar results have already been proven in [52, 84] for the periodic case.

3.5.2 Fourier polynomials on full grids and hyperbolic crosses

In the last subsection, we examined the so-called h-version of sparse grids. This means that the degree of the piecewise polynomials is fixed and their support is refined with increasing level. In the following, we instead consider a type of spectral/p-version ap-

proach, where we increase the maximum frequency of global Fourier polynomials. In particular, we are interested in polynomials on hyperbolic crosses. Throughout this subsection, let $T = (-\pi, \pi)^m$. We identify opposite hyperplanes, i.e. we deal with functions which are 2π-periodic in every coordinate.

Periodic Sobolev spaces

Since we deal with complex-valued Fourier polynomials in this subsection, our corresponding function spaces also have to be complex-valued. If we restrict the Sobolev spaces $H^s(T; \mathbb{C}^d)$ and $H^s_{\text{mix}}(T; \mathbb{C}^d)$ to periodic functions, we obtain the periodic Sobolev spaces on $\bar{H}^s(T; \mathbb{C}^d)$ and $\bar{H}^s_{\text{mix}}(T; \mathbb{C}^d)$. For convenience, we also make a slight change of the norms for these periodic Sobolev spaces compared to their non-periodic originals. The reason is that we can consider Fourier series instead of Fourier transformations for periodic functions. To this end, we set

$$\|f\|_{\bar{H}^s(T;\mathbb{C}^d)} := \left\| \sum_{\mathbf{k}\in\mathbb{Z}^m} c_{\mathbf{k}}(f)(1 + \|\mathbf{k}\|_{\ell_2}^2)^{\frac{s}{2}} e^{i\mathbf{k}^T\mathbf{t}} \right\|_{L_2(T;\mathbb{C}^d)}$$

and

$$\|f\|_{\bar{H}^s_{\text{mix}}(T;\mathbb{C}^d)} := \left\| \sum_{\mathbf{k}\in\mathbb{Z}^m} c_{\mathbf{k}}(f) \prod_{j=1}^{m} (1 + |k_j|^2)^{\frac{s}{2}} e^{i\mathbf{k}^T\mathbf{t}} \right\|_{L_2(T;\mathbb{C}^d)},$$

where $\mathbf{t} \in T$ is the spatial variable and

$$c_{\mathbf{k}}(f) := \frac{1}{(2\pi)^m} \mathcal{F}(f)(\mathbf{k}) = \frac{1}{(2\pi)^m} \int_T f(\mathbf{t}) e^{-i\mathbf{k}^T\mathbf{t}} \mathrm{d}\mathbf{t}$$

denotes the $\mathbf{k}$-th Fourier coefficient. For more details on these norms, we refer to [52, 84].

Full grids and hyperbolic crosses

The space of trigonometric/Fourier polynomials on a full grid of level $k \in \mathbb{N}$ is defined similarly to the full grid space for piecewise linear functions, which we introduced in the previous subsection. To this end, let

$$\mathcal{T}_k^{\text{full},d} := \text{span}\left\{ \exp\left(i\mathbf{l}^T\mathbf{t}\right) \cdot \mathbf{e}_j \mid \mathbf{l} \in \mathbb{Z}^m, |\mathbf{l}|_\infty \leq 2^k \text{ and } j = 1, \ldots, d \right\}, \tag{3.43}$$

where i is the imaginary unit. Another way to interpret this is that $\mathcal{T}_k^{\text{full},d}$ contains all those periodic L_2 functions for which $c_{\mathbf{l}}(f) = \mathbf{0} \in \mathbb{C}^d$ for all $\mathbf{l} \in \mathbb{Z}^m$ with $|\mathbf{l}|_\infty > 2^k$. Following [78], the space of Fourier polynomials on the so-called hyperbolic cross of level

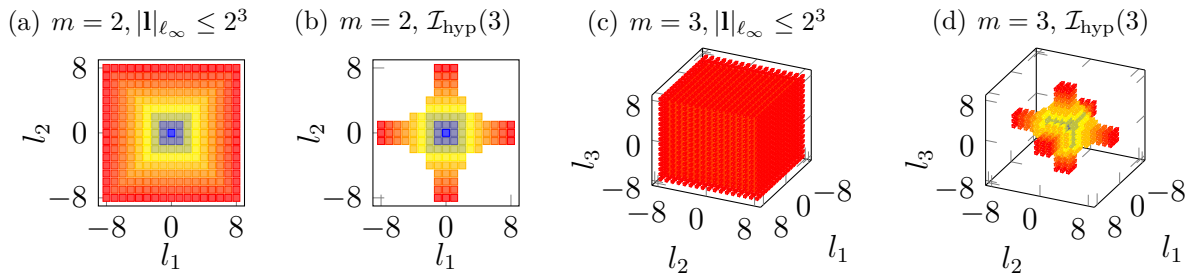

Fig. 3.3: Full grid and hyperbolic cross index sets for $k = 3$ in dimensions $m = 2, 3$. We set $\mathcal{I}_{\text{hyp}}(k) := \{\mathbf{l} \in \mathbb{Z}^m \mid \sum_{n=1}^m \log_2(\max(|l_n|, 1)) \leq k\}$

$k \in \mathbb{N}$ can be written as

$$\mathcal{T}_k^{\text{hyp},d} := \text{span}\left\{\exp\left(i\mathbf{l}^T\mathbf{t}\right) \cdot \mathbf{e}_j \mid \mathbf{l} \in \mathbb{Z}^m, \prod_{n=1}^m (\max(|l_n|, 1)) \leq 2^k\right\}. \tag{3.44}$$

Note that the inequality

$$\prod_{n=1}^m (\max(|l_n|, 1)) \leq 2^k \Leftrightarrow \sum_{n=1}^m \log_2(\max(|l_n|, 1)) \leq k,$$

which defines the multilevel indices contained in the hyperbolic cross, can be interpreted as an analogue to the condition $\zeta_m(\mathbf{l}) \leq k$ for sparse grids. An illustration of the hyperbolic cross index sets in contrast to full grid index sets can be found in figure 3.3. As we already mentioned for the vector-valued sparse grid spaces, we could also define the Fourier polynomial spaces for functions with image in any separable Banach space E by substituting $\mathbf{e}_j$ by a suitable basis.

We directly observe that the dimension of the full grid space is

$$\dim(\mathcal{T}_k^{\text{full},d}) = d \cdot \mathcal{O}\left(2^{km}\right).$$

The number of degrees of freedom of the hyperbolic cross space can be bounded from above by

$$\dim(\mathcal{T}_k^{\text{hyp},d}) = d \cdot \mathcal{O}\left(2^k k^{m-1}\right), \tag{3.45}$$

see [78]. This is analogous to the sparse grid space $\mathcal{V}_k^{\text{sparse},d}$.

Norm equivalences and inverse inequalities

To obtain norm equivalences similar to (3.37) and (3.38), we first need to introduce a decomposition of the spaces $\mathcal{T}_k^{\text{full},d}$ and $\mathcal{T}_k^{\text{hyp},d}$. To this end, let $\text{Par}_0 := \{-1, 0, 1\}$ and

let $\mathrm{Par}_l := \{z \in \mathbb{Z} \mid 2^{l-1} < |z| \leq 2^l\}$ for all $l \in \mathbb{N}^+ = \mathbb{N} \setminus \{0\}$. Let furthermore

$$\mathrm{Par}_{\mathbf{l}} := \mathrm{Par}_{l_1} \times \ldots \times \mathrm{Par}_{l_m}$$

for $\mathbf{l} \in \mathbb{N}^m$. Then, we can decompose a periodic function $f \in L_2(T; \mathbb{C}^d)$ by $f = \sum_{\mathbf{l} \in \mathbb{N}^m} \bar{w}_{\mathbf{l}}$ with

$$\bar{w}_{\mathbf{l}}(\mathbf{t}) = \sum_{\mathbf{k} \in \mathrm{Par}_{\mathbf{l}}} c_{\mathbf{k}}(f) e^{-i\mathbf{k}^T \mathbf{t}}.$$

This decomposition, together with Parseval's identity, gives us the desired norm equivalences for the periodic Sobolev spaces:

$$\|f\|^2_{\bar{H}^s(T;\mathbb{C}^d)} \simeq \sum_{\mathbf{l} \in \mathbb{N}^m} 2^{2s|\mathbf{l}|_{\ell_\infty}} \|\bar{w}_{\mathbf{l}}\|^2_{L_2(T;\mathbb{C}^d)} = \sum_{\mathbf{l} \in \mathbb{N}^m} 2^{2s|\mathbf{l}|_{\ell_\infty}} \sum_{\mathbf{k} \in \mathrm{Par}_{\mathbf{l}}} \|c_{\mathbf{k}}(f)\|^2_{\ell_2} \quad \forall\, s \geq 0 \tag{3.46}$$

and

$$\|f\|^2_{\bar{H}^s_{\mathrm{mix}}(T;\mathbb{C}^d)} \simeq \sum_{\mathbf{l} \in \mathbb{N}^m} 2^{2s|\mathbf{l}|_{\ell_1}} \|\bar{w}_{\mathbf{l}}\|^2_{L_2(T;\mathbb{C}^d)} = \sum_{\mathbf{l} \in \mathbb{N}^m} 2^{2s|\mathbf{l}|_{\ell_1}} \sum_{\mathbf{k} \in \mathrm{Par}_{\mathbf{l}}} \|c_{\mathbf{k}}(f)\|^2_{\ell_2} \quad \forall\, s \geq 0. \tag{3.47}$$

For details on these results, see [84]. Taking the definition of $\bar{w}_{\mathbf{l}}$ into account, we observe that $\bar{w}_{\mathbf{l}} = 0$ for all $\mathbf{l} \in \mathbb{N}^m$ with

$$2^{|\mathbf{l}|_{\ell_\infty}} > 2^k \;\Leftrightarrow\; |\mathbf{l}|_{\ell_\infty} > k$$

if $f \in \mathcal{T}_k^{\mathrm{full},d}$. Analogously, we obtain $\bar{w}_{\mathbf{l}} = 0$ for all $\mathbf{l} \in \mathbb{N}^m$ with

$$\sum_{n=1}^{m} \log_2(\max(2^{l_n}, 1)) > k \;\Leftrightarrow\; |\mathbf{l}|_{\ell_1} > k$$

if $f \in \mathcal{T}_k^{\mathrm{hyp},d}$. With the above results, we can again easily derive the inverse, Bernstein-type inequalities for $s \geq 0$. We have

$$\|f\|^2_{\bar{H}^s(T;\mathbb{C}^d)} \overset{(3.46)}{\simeq} \sum_{|\mathbf{l}|_{\ell_\infty} \leq k} 2^{2s|\mathbf{l}|_{\ell_\infty}} \|w_{\mathbf{l}}\|^2_{L_2(T;\mathbb{C}^d)} \lesssim 2^{2sk} \sum_{|\mathbf{l}|_{\ell_\infty} \leq k} \|w_{\mathbf{l}}\|^2_{L_2(T;\mathbb{C}^d)} \overset{(3.46)}{\simeq} 2^{2sk} \|f\|^2_{L_2(T;\mathbb{C}^d)} \tag{3.48}$$

for $f \in \mathcal{T}_k^{\mathrm{full},d}$ and

$$\|f\|^2_{\bar{H}^s_{\mathrm{mix}}(T;\mathbb{C}^d)} \overset{(3.47)}{\simeq} \sum_{|\mathbf{l}|_{\ell_1} \leq k} 2^{2s|\mathbf{l}|_{\ell_1}} \|w_{\mathbf{l}}\|^2_{L_2(T;\mathbb{C}^d)} \lesssim 2^{2sk} \sum_{|\mathbf{l}|_{\ell_1} \leq k} \|w_{\mathbf{l}}\|^2_{L_2(T;\mathbb{C}^d)} \overset{(3.47)}{\simeq} 2^{2sk} \|f\|^2_{L_2(T;\mathbb{C}^d)} \tag{3.49}$$

for $f \in \mathcal{T}_k^{\mathrm{hyp},d}$. For more inverse inequalities of spaces of Fourier polynomials, we refer to [78].

Best approximation error

The rate of the L_2 best approximation error behaves similarly as for the piecewise linear prewavelets. However, due to the usage of the Fourier polynomials, we can exploit additional smoothness of the approximated function and derive results also for Sobolev smoothness $s > 2$. In the full grid case, we obtain

$$\inf_{f \in \mathcal{T}_k^{\text{full},d}} \|f - g\|^2_{L_2(T;\mathbb{C}^d)} \lesssim d \cdot 2^{-2sk} \|g\|^2_{\bar{H}^s(T;\mathbb{C}^d)} \quad \text{if } g \in \bar{H}^s(T;\mathbb{C}^d) \text{ with } s > 0. \tag{3.50}$$

The best approximation error for the hyperbolic cross behaves like

$$\inf_{f \in \mathcal{T}_k^{\text{hyp},d}} \|f - g\|^2_{L_2(T;\mathbb{C}^d)} \lesssim d \cdot 2^{-2sk} \|g\|^2_{\bar{H}^s_{\text{mix}}(T;\mathbb{C}^d)} \quad \text{if } g \in \bar{H}^s_{\text{mix}}(T;\mathbb{C}^d) \text{ with } s > 0. \tag{3.51}$$

Here, the $\lesssim$ constants only depend on s and m. For the proofs, we refer to theorems II.3.2 and III.3.2 of [78]. Note, to this end, that the function classes $\mathbf{SW}_2^s$ and $\mathbf{MW}_2^s$ considered in [78] correspond to the unit balls of our spaces $\bar{H}^s(T)$ and $\bar{H}^s_{\text{mix}}(T)$. For details on this equivalence, we refer to theorem 2.7 of [84].

4 Constrained regression

The problem of regression has a long history in mathematics going back to the first years of the 19th century, when the method of least-squares regression was introduced by Legendre [56] and Gauß [37]. However, the name "regression" was unconventional until 1886, when Galton described the regression of the average height of an offspring with respect to the height of its tall parents, see [34]. Ever since, the term "regression analysis" described the estimation of the relationship between two or more variables.

Due to many different underlying model assumptions, a vast amount of different methods for regression analysis exist. The model is usually fitted to a specific field of application or to the computational infrastructure at hand. The resulting algorithms range from simple linear function regression to complex non-linear Bayesian maximum a-posteriori estimators, support vector machines and artificial neural networks, see [17, 26, 46] for an overview.

In this chapter, we focus on constrained vector-valued least-squares regression. Here, the term constrained refers to the fact that we only consider functions which fulfill a certain norm bound. The problem of finding the optimal regression function translates to a quadratic functional minimization problem over a fixed search set. Due to its benign nature, error estimates and convergence rates for the corresponding algorithm already exist for many choices of the search set, see e.g. [23, 46]. However, for the case of constrained regression in a finite-dimensional search space, the known theoretical results cannot be applied to obtain a non-trivial upper bound on the regression error. Therefore, we introduce a technique relying on Jackson- and Bernstein-type inequalities to treat the finite-dimensional setting. Our result can then be applied to any finite-dimensional constrained regression problem, such as sparse grid-based regression, see [11, 15], for example.

This chapter is structured as follows: In section 4.1, we discuss the general framework of vector-valued regression and focus on the common case of least-squares regression in Bochner spaces. In section 4.2, we investigate the solvability of the constrained regression problem and present a splitting of the overall error into the *bias* and the *sampling error* term. We proceed with well-known techniques from interpolation theory to derive upper bounds on the bias in section 4.3. Furthermore, we deal with the finite-dimensional case separately and establish bounds based on given Jackson and Bernstein estimates. In section 4.4, we review the concept of covering numbers to determine probabilistic bounds for the sampling error. We apply the results from sections 4.3 and 4.4 to obtain estimates on the overall regression error for several examples in section 4.5. Finally, we conclude this chapter with a short summary in section 4.6. To give a brief overview on

the most relevant functions and spaces/sets in this chapter, we included table 4.1.

4.1 The regression functional

In this section, we introduce the general vector-valued regression problem. We define the so-called regression functional and formulate our task as a minimization problem over a suitable set of functions. Let us emphasize that we deal with multivariate regression of vector-valued functions and, thus, we have to carefully distinguish between the dimension $m \in \mathbb{N}$ of the domain and the dimension $d \in \mathbb{N} \cup \{\infty\}$ of the image of a vector-valued function. The proofs in this section are essentially adjusted versions of the proofs in [23], where the scalar-valued case $d = 1$ is treated.

Regression with given measure

Let $T \subset \mathbb{R}^m$ be a bounded, open domain and let $(E, \langle \cdot, \cdot \rangle_E)$ be a separable real Hilbert space[1]. Let $\Sigma(T \times E)$ be the smallest σ-algebra which contains

$$\{A \times B \mid A \in \Sigma_{\mathcal{L}}(T), B \in \Sigma_{\mathcal{B}}(E)\},$$

i.e. $\Sigma(T \times E)$ is the product algebra of the Lebesgue algebra $\Sigma_{\mathcal{L}}(T)$ on T and the Borel algebra $\Sigma_{\mathcal{B}}(E)$ on E. Let ρ be a probability measure on $T \times E$ with respect to $\Sigma(T \times E)$ such that the marginal measure

$$\rho_T(\cdot) := \rho(\cdot, E)$$

on T is a non-degenerate probability measure with respect to $\Sigma_{\mathcal{L}}(T)$, i.e. $\rho_T(T) = 1$ and

$$\rho_T(U) > 0 \tag{4.1}$$

for each non-empty open set $U \in \Sigma_{\mathcal{L}}(T)$. Moreover, we assume that the conditional measures $\rho(\cdot|\mathbf{t})$ with respect to $\mathbf{t} \in T$, which fulfill

$$\int_{T \times E} G(\mathbf{t}, \mathbf{x}) \, \mathrm{d}\rho(\mathbf{t}, \mathbf{x}) = \int_T \left(\int_E G(\mathbf{t}, \mathbf{x}) \, \mathrm{d}\rho(\mathbf{x}|\mathbf{t}) \right) \, \mathrm{d}\rho_T(\mathbf{t}) \tag{4.2}$$

for all integrable functions $G : T \times E \to \mathbb{R}$, are probability measures on $\Sigma_{\mathcal{B}}(E)$. A general multivariate vector-valued regression problem reads

$$\text{Find } \hat{f} := \operatorname*{arg\,min}_{f \in L_{2,\rho_T}(T;E)} \mathcal{E}(f) \text{ with } \mathcal{E}(f) := \int_{T \times E} \psi(f(\mathbf{t}), \mathbf{x}) \, \mathrm{d}\rho(\mathbf{t}, \mathbf{x}). \tag{A}$$

[1] Note that some of our results also hold in the more general case where E is a Banach space. However, since the most important statements are only valid if E is a Hilbert space, we will focus on this case only.

Table 4.1: Overview on relevant functions, sets and variables for constrained regression.

$T \subset \mathbb{R}^m$	open and bounded domain
$(E, \langle \cdot, \cdot \rangle_E)$	separable Hilbert space
$\rho : \Sigma(T \times E) \to [0,1]$	probability measure on the product algebra of $\Sigma_{\mathcal{L}}(T)$ and $\Sigma_{\mathcal{B}}(E)$
$\rho_T : \Sigma_{\mathcal{L}}(T) \to [0,1]$	marginal measure of ρ with respect to the first coordinate
$L_{2,\rho_T}(T;E)$	L_2 Bochner space (with respect to ρ_T) of functions with domain T and image in E
$C(T;E)$	Banach space of continuous functions from T to E equipped with the sup-norm
$\mathcal{H} \subset C(T;E)$	search space, a reflexive Banach space continuously embedded into $C(T;E)$
$\mathcal{H}_b \subset \mathcal{H}$	search set, a closed ball of radius b in $\mathcal{H}$ centered at 0 w.r.t. the norm $\|\cdot\|_{\mathcal{H}}$
$V_k \subset C(T;E)$	finite-dimensional search space, continuously embedded into $C(T;E)$
$V_{k,b} \subset V_k$	finite-dimensional search set, a closed ball of radius b in V_k centered at 0 w.r.t. the norm $\|\cdot\|_{V_k}$
$r \in (0,\infty)$	radius of a bounding ball of the support of $\rho(T,\cdot)$
$\mathcal{Z}_n \in (T \times E)^n$	n i.i.d. samples $(\mathbf{t}_i, \mathbf{x}_i)_{i=1}^n$ drawn according to ρ
$N_k \in \mathbb{N}$	degrees of freedom of the search space, $N_k = \dim(V_k)$
$\psi : E \times E \to [0,\infty)$	cost function, measures how close the arguments are to each other
$M_\psi \in (0,\infty)$	upper bound on $\psi(f(\mathbf{t}),\mathbf{x})$ for all f in the search set and almost every $\mathbf{t} \in T, \mathbf{x} \in E$
$\mathcal{E} : L_{2,\rho_T}(T;E) \to [0,\infty]$	the target functional for the regression problem
$\mathcal{E}_{\mathcal{Z}_n} : L_{2,\rho_T}(T;E) \to [0,\infty]$	the target functional for the finite sample regression problem
$\hat{f} \in L_{2,\rho_T}(T;E)$	minimizer of $\mathcal{E}$ in $L_{2,\rho_T}(T;E)$
$f_\rho \in L_{2,\rho_T}(T;E)$	$\hat{f}$ for the special case $\psi(\mathbf{x},\mathbf{y}) = \|\mathbf{x}-\mathbf{y}\|_E^2$
$f_X \in X \subset L_{2,\rho_T}(T;E)$	minimizer of $\mathcal{E}$ over X
$f_{\mathcal{Z}_n,X} \in X \subset L_{2,\rho_T}(T;E)$	minimizer of $\mathcal{E}_{\mathcal{Z}_n}$ over X

The function $\psi : E \times E \to [0, \infty)$ is called *cost function*. It penalizes large distances between the two input vectors and it fulfills $\psi(\mathbf{x}, \mathbf{x}) = 0$ for every $\mathbf{x} \in E$. An example for ψ would be a metric, however many commonly used cost functions are no metrics. The so-called *ϵ-insensitive loss* function

$$\psi(\mathbf{x}, \mathbf{y}) = \begin{cases} 0 & \text{if } \|\mathbf{x} - \mathbf{y}\|_E \leq \epsilon, \\ \|\mathbf{x} - \mathbf{y}\|_E - \epsilon & \text{else} \end{cases},$$

for instance is a well-known example which does not fulfill the properties of a metric since the triangle inequality does not hold, see [74, 85] for details.

We will solely focus on the squared distance

$$\psi(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} - \mathbf{y}\|_E^2 \tag{4.3}$$

for most parts of this thesis because of its strict convexity and its relation to the $L_{2,\rho_T}(T; E)$ Bochner norm, which we will explore in the following. Nonetheless, many results - such as the existence and the uniqueness of minimizers - can easily be extended to a more general type of cost functions.

Definition 4.1 [Qualified cost function]
A positive cost function $\psi : E \times E \to [0, \infty)$ which fulfills

$$\psi(\mathbf{x}, \mathbf{y}) = \tilde{\psi}(\mathbf{x} - \mathbf{y}) \tag{4.4}$$

with even, continuous and convex $\tilde{\psi}$ which fulfills $\tilde{\psi}(\mathbf{0}) = 0$ is called ***qualified cost function****.*

Trivially, $\psi(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} - \mathbf{y}\|_E^2$ from (4.3) and the ϵ-insensitive loss function are qualified cost functions. For our further analysis, we assume that there exists an $r > 0$ such that[2]

$$\rho(T \times \overline{U_r(\mathbf{0})}) = 1,$$

where $\overline{U_r(\mathbf{0})}$ denotes the closed ball of radius r in E with center $\mathbf{0}$, i.e.

$$\|\mathbf{x}\|_E \leq r \quad \text{for } \rho\text{-almost every } (\mathbf{t}, \mathbf{x}) \in T \times E. \tag{4.5}$$

[2]Note that it would be sufficient if the milder assumption

$$\int_E \|\mathbf{x}\|_E \, \mathrm{d}\rho(\mathbf{x}|\mathbf{t}) \leq r < \infty$$

held for ρ_T-almost every $\mathbf{t} \in T$ in order to derive most of our results. However, when considering the sampling error in section 4.4, the almost surely bound of $\|\mathbf{x}\|_E$ is an essential requirement to obtain the necessary uniform bound on the cost function (4.3).

Regression with given samples

In real-world applications, the measure ρ is unknown and we only have access to a finite set $\mathcal{Z}_n$ of n sample points

$$\mathcal{Z}_n := ((\mathbf{t}_i, \mathbf{x}_i))_{i=1}^n \in (T \times E)^n \tag{4.6}$$

which we assume to be drawn independently of each other and to be distributed according to ρ. This means that $\mathcal{Z}_n \sim \rho^n$. For a given sample set $\mathcal{Z}_n$, we define the empirical measures on $T \times E$ and T, respectively, by

$$\delta_{\mathcal{Z}_n} := \frac{1}{n} \sum_{i=1}^n \delta_{\mathbf{t}_i, \mathbf{x}_i} \quad \text{and} \quad \delta_{\mathbf{t}_1, \dots, \mathbf{t}_n} := \frac{1}{n} \sum_{i=1}^n \delta_{\mathbf{t}_i}.$$

Here, $\delta_{\mathbf{t}_i, \mathbf{x}_i}$ is the Dirac measure centered in $(\mathbf{t}_i, \mathbf{x}_i) \in T \times E$ and $\delta_{\mathbf{t}_i}$ is the Dirac measure centered in $\mathbf{t}_i \in T$. Instead of considering (A) directly, we now substitute ρ by $\delta_{\mathcal{Z}_n}$ to obtain the regression problem for the finite sample set $\mathcal{Z}_n$:

$$\text{Find} \quad \operatorname*{arg\,min}_{f \in L_{2,\rho_T}(T;E)} \mathcal{E}_{\mathcal{Z}_n}(f) \quad \text{with } \mathcal{E}_{\mathcal{Z}_n}(f) := \frac{1}{n} \sum_{i=1}^n \psi(f(\mathbf{t}_i), \mathbf{x}_i). \tag{B}$$

Note that the point evaluation $f(\mathbf{t}_i)$ is not necessarily defined for L_{2,ρ_T} functions. However, when we are dealing with continuous functions later on, we automatically circumvent this problem.

Properties of $\mathcal{E}$ and $\mathcal{E}_{\mathcal{Z}_n}$

First, let us consider the function f_ρ, which will be of importance when analyzing the error made by a regression algorithm.

Lemma 4.2
Let $f_\rho : T \to E$ be defined by

$$f_\rho(\mathbf{t}) := \int_E \mathbf{x} \, \mathrm{d}\rho(\mathbf{x}|\mathbf{t}). \tag{4.7}$$

Then $f_\rho \in L_{\infty,\rho_T}(T;E) \subset L_{2,\rho_T}(T;E)$.

Proof. Applying the norm inequality (3.6) for Bochner integrals, we obtain

$$\begin{aligned} \|f_\rho\|_{L_{\infty,\rho_T}(T;E)} &= \operatorname*{ess\,sup}_{\mathbf{t} \in T} \left\| \int_E \mathbf{x} \, \mathrm{d}\rho(\mathbf{x}|\mathbf{t}) \right\|_E \leq \operatorname*{ess\,sup}_{\mathbf{t} \in T} \int_E \|\mathbf{x}\|_E \, \mathrm{d}\rho(\mathbf{x}|\mathbf{t}) \\ &\overset{(4.5)}{\leq} \operatorname*{ess\,sup}_{\mathbf{t} \in T} r = r < \infty. \end{aligned}$$

Since ρ_T is a probability measure, $L_{\infty,\rho_T}(T;E) \subset L_{2,\rho_T}(T;E)$ holds, see (3.5). □

For the further analysis of the behavior of $\mathcal{E}$ and $\mathcal{E}_{\mathcal{Z}_n}$, the Lipschitz regularity of these error functionals will be crucial. Therefore, we provide the following lemma for the quadratic distance function (4.3), see also [23].

Lemma 4.3
Let $f_1, f_2 \in L_{\infty,\rho_T}(T;E)$ with $\|f_i\|_{L_{\infty,\rho_T}(T;E)} \leq M$ for $i = 1, 2$. Let furthermore $\psi(\mathbf{x},\mathbf{y}) = \|\mathbf{x}-\mathbf{y}\|_E^2$. Then, $\mathcal{E}$ and $\mathcal{E}_{\mathcal{Z}_n}$ fulfill the Lipschitz conditions

$$|\mathcal{E}(f_1) - \mathcal{E}(f_2)| \leq C\|f_1 - f_2\|_{L_{1,\rho_T}(T;E)} \leq C\|f_1 - f_2\|_{L_{2,\rho_T}(T;E)} \tag{4.8}$$

and

$$|\mathcal{E}_{\mathcal{Z}_n}(f_1) - \mathcal{E}_{\mathcal{Z}_n}(f_2)| \leq C\|f_1 - f_2\|_{L_{1,\delta_{\mathbf{t}_1,\ldots,\mathbf{t}_n}}(T;E)} \leq C\|f_1 - f_2\|_{L_{2,\delta_{\mathbf{t}_1,\ldots,\mathbf{t}_n}}(T;E)} \tag{4.9}$$

with $C := 2(M + r)$, where r stems from (4.5).

Proof. Since E is a Hilbert space, we obtain

$$\begin{aligned} |\psi(f_1(\mathbf{t}),\mathbf{x}) - \psi(f_2(\mathbf{t}),\mathbf{x})| &= \|f_1(\mathbf{t}) - \mathbf{x}\|_E^2 - \|f_2(\mathbf{t}) - \mathbf{x}\|_E^2 \\ &= \langle f_1(\mathbf{t}) - f_2(\mathbf{t}), f_1(\mathbf{t}) + f_2(\mathbf{t}) - 2\mathbf{x}\rangle_E \\ &\leq \|f_1(\mathbf{t}) - f_2(\mathbf{t})\|_E \cdot (\|f_1(\mathbf{t})\|_E + \|f_2(\mathbf{t})\|_E + 2\|\mathbf{x}\|_E) \\ &\leq 2(M + r)\|f_1(\mathbf{t}) - f_2(\mathbf{t})\|_E \end{aligned} \tag{4.10}$$

for almost every $(\mathbf{t},\mathbf{x}) \in T \times E$. With this, we obtain

$$\begin{aligned} |\mathcal{E}(f_1) - \mathcal{E}(f_2)| &\leq \int_{T\times E} |\psi(f_1(\mathbf{t}),\mathbf{x}) - \psi(f_2(\mathbf{t}),\mathbf{x})| \,\mathrm{d}\rho(\mathbf{t},\mathbf{x}) \\ &\leq C\int_T \|f_1(\mathbf{t}) - f_2(\mathbf{t})\|_E \int_E 1 \,\mathrm{d}\rho(\mathbf{x}|\mathbf{t}) \,\mathrm{d}\rho_T(\mathbf{t}) \\ &= C\int_T \|f_1(\mathbf{t}) - f_2(\mathbf{t})\|_E \,\mathrm{d}\rho_T(\mathbf{t}) \\ &= C\|f_1 - f_2\|_{L_{1,\rho_T}(T;E)}. \end{aligned}$$

From (3.8), we obtain $\|f_1 - f_2\|_{L_{1,\rho_T}(T;E)} \leq \|f_1 - f_2\|_{L_{2,\rho_T}(T;E)}$ for all $f_1 - f_2 \in L_{2,\rho_T}(T;E)$, which finally shows (4.8).

The proof for the inequality (4.9) works analogously: We have

$$\begin{aligned} |\mathcal{E}_{\mathcal{Z}_n}(f_1) - \mathcal{E}_{\mathcal{Z}_n}(f_2)| &\leq \frac{1}{n}\sum_{i=1}^n |\psi(f_1(\mathbf{t}_i),\mathbf{x}_i) - \psi(f_2(\mathbf{t}_i),\mathbf{x}_i)| \\ &\leq \frac{C}{n}\sum_{i=1}^n \|f_1(\mathbf{t}_i) - f_2(\mathbf{t}_i)\|_E = C\|f_1 - f_2\|_{L_{1,\delta_{\mathbf{t}_1,\ldots,\mathbf{t}_n}}(T;E)}. \end{aligned}$$

Again, applying (3.8) to $\|f_1 - f_2\|_{L_{1,\delta_{\mathbf{t}_1,\ldots,\mathbf{t}_n}}(T;E)}$ gives the final result (4.9). □

We conclude this subsection with an important property of the functional $\mathcal{E}$ and the function f_ρ.

Lemma 4.4
Let $f \in L_{2,\rho_T}(T;E)$ and let $\psi(\mathbf{x},\mathbf{y}) = \|\mathbf{x}-\mathbf{y}\|_E^2$, see (4.3). Then, we obtain

$$\mathcal{E}(f) - \mathcal{E}(f_\rho) = \|f - f_\rho\|^2_{L_{2,\rho_T}(T;E)} \tag{4.11}$$

for f_ρ from (4.7).

Proof. It holds

$$\begin{aligned}
\mathcal{E}(f) &= \int_{T\times E} \|f(\mathbf{t}) - f_\rho(\mathbf{t}) + f_\rho(\mathbf{t}) - \mathbf{x}\|_E^2 \,\mathrm{d}\rho(\mathbf{t},\mathbf{x}) \\
&= \int_{T\times E} \|f(\mathbf{t}) - f_\rho(\mathbf{t})\|_E^2 \,\mathrm{d}\rho(\mathbf{t},\mathbf{x}) + \int_{T\times E} \|f_\rho(\mathbf{t}) - \mathbf{x}\|_E^2 \,\mathrm{d}\rho(\mathbf{t},\mathbf{x}) \\
&+ \int_{T\times E} 2\langle f(\mathbf{t}) - f_\rho(\mathbf{t}), f_\rho(\mathbf{t}) - \mathbf{x}\rangle_E \,\mathrm{d}\rho(\mathbf{t},\mathbf{x}) \\
&= \|f - f_\rho\|^2_{L_{2,\rho_T}(T;E)} + \mathcal{E}(f_\rho) + 2\int_{T\times E} \langle f(\mathbf{t}) - f_\rho(\mathbf{t}), f_\rho(\mathbf{t}) - \mathbf{x}\rangle_E \,\mathrm{d}\rho(\mathbf{t},\mathbf{x})
\end{aligned}$$

Note that the last summand is zero because of the definition of f_ρ. Indeed, we have

$$\begin{aligned}
&\int_{T\times E} \langle f(\mathbf{t}) - f_\rho(\mathbf{t}), f_\rho(\mathbf{t}) - \mathbf{x}\rangle_E \,\mathrm{d}\rho(\mathbf{t},\mathbf{x}) \\
&= \int_T \int_E \langle f(\mathbf{t}) - f_\rho(\mathbf{t}), f_\rho(\mathbf{t}) - \mathbf{x}\rangle_E \,\mathrm{d}\rho(\mathbf{x}|\mathbf{t}) \,\mathrm{d}\rho_T(\mathbf{t}) \\
&= \int_T \langle f(\mathbf{t}) - f_\rho(\mathbf{t}), f_\rho(\mathbf{t}) - \int_E \mathbf{x} \,\mathrm{d}\rho(\mathbf{x}|\mathbf{t})\rangle_E \,\mathrm{d}\rho_T(\mathbf{t}) \\
&= \int_T \langle f(\mathbf{t}) - f_\rho(\mathbf{t}), f_\rho(\mathbf{t}) - f_\rho(\mathbf{t})\rangle_E \,\mathrm{d}\rho_T(\mathbf{t}) = 0,
\end{aligned}$$

where we used the commutativity of the Bochner integral and the scalar product, see (3.7). Therefore, (4.11) is proven. □

As a direct consequence of lemma 4.4, we obtain that f_ρ is the unique solution to (A) for the cost function (4.3).

Summary

We shortly summarize the results from this section:

- Our goal is to minimize (A). However, usually the measure ρ is unknown and the evaluation of $\mathcal{E}$ is not possible. Therefore, we rely on a finite sample variant of the minimization problem, namely (B).

- The error which is made by computing an approximation f to $\hat{f}$ from (A) is given by $\mathcal{E}(f) - \mathcal{E}(\hat{f})$.

- For the cost function $\psi(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} - \mathbf{y}\|_E^2$, we obtain that $\hat{f} = f_\rho$ is the unique minimizer of the regression problem (A). Lemma 4.4 shows us that the overall error can be written as

$$\mathcal{E}(f) - \mathcal{E}(f_\rho) = \|f - f_\rho\|^2_{L_{2,\rho_T}(T;E)}.$$

  This relation allows us to employ bounds on the $L_{2,\rho_T}(T; E)$ approximation error to estimate the regression error.

4.2 Solutions to the regression problem and the overall error

In this section, we discuss the overall error that is made by a regression algorithm in terms of the value of the functional $\mathcal{E}$, i.e. the difference $\mathcal{E}(f)-\mathcal{E}(\hat{f})$ for an (approximate) solution f to the regression problem. First, we need to discuss under which conditions the problems (A) and (B) are well-defined in the sense that there exists a (unique) minimizer. Subsequently, we have a look at the overall regression error in more detail.

In (A), we considered the original task of solving the minimization problem for $f \in L_{2,\rho_T}(T; E)$. However, since we deal with point evaluations of functions in (B), it makes sense to restrict our search to a set which consists of point evaluable functions. To this end, we consider a Banach space $(\mathcal{H}, \|\cdot\|_\mathcal{H})$, which we refer to as *search space* from now on. We assume that the search space is continuously embedded into the space $C(T; E)$ of vector-valued continuous functions equipped with the maximum norm

$$\|f\|_\infty := \sup_{\mathbf{t}\in T} \|f(t)\|_E. \tag{4.12}$$

The most important candidates for such a space $\mathcal{H}$ are vector-valued reproducing kernel Hilbert spaces, which we studied in section 3.3. Note however, that also more general Banach spaces can be considered here. Usually, we restrict the minimization problem (A) or (B), respectively, to a subset of the search space, e.g. a bounded ball in $\mathcal{H}$, which we refer to as *search set*.

We study the existence and uniqueness of minimizers over bounded balls in $\mathcal{H}$ in subsection 4.2.1. Subsequently, we introduce the overall error for regression over the search set in subsection 4.2.2, where we also consider a decomposition of this error into the so-called bias part and the sampling error part.

4.2.1 Existence and uniqueness of minimizers

In section 4.1, we defined the functional $\mathcal{E}_{\mathcal{Z}_n}$ by substituting the unknown measure ρ by independent, identically distributed samples $\mathcal{Z}_n$, see also (4.6). Subsequently, we obtained the finite sample regression problem (B). The minimization of (B) is obviously ill-posed in the sense that infinitely many minimizers exist if the whole space $L_{2,\rho_T}(T;E)$ is taken as search set. Therefore, it makes sense to consider a more restrictive search set. At the beginning of this section we already motivated why it makes sense to deal with functions which are at least continuous. We now go into more detail and discuss which search sets are suitable to obtain a finite sample minimization problem which is uniquely solvable. To this end, we study the case where the search set is a bounded ball in a specific Banach space $\mathcal{H}$ instead of the whole space $L_{2,\rho_T}(T;E)$. Most concepts introduced in this section are based on the ideas in [22],[23], where the scalar-valued case $E = \mathbb{R}$ is treated.

Bounded search sets

Let $(\mathcal{H}, \|\cdot\|_{\mathcal{H}})$ be a reflexive, real Banach space of functions mapping from T to E. We assume that $\mathcal{H}$ is continuously embedded into $(C(T;E), \|\cdot\|_\infty)$, where the $\|\cdot\|_\infty$ norm is defined as in (4.12). We denote by

$$c_{\mathcal{H}} := \|\mathrm{id} : \mathcal{H} \hookrightarrow C(T;E)\|_{\mathcal{L}(\mathcal{H},C(T;E))} \tag{4.13}$$

the corresponding embedding constant, where $\|\cdot\|_{\mathcal{L}(\mathcal{H},C(T;E))}$ is the standard norm for linear operators, see (3.23). Since ρ_T is a probability measure, the embedding from $C(T;E)$ into $L_{2,\rho_T}(T;E)$ is naturally continuous with embedding constant 1, see also (3.8). Therefore, the resulting chain of continuous embeddings can be written as

$$(\mathcal{H}, \|\cdot\|_{\mathcal{H}}) \hookrightarrow (C(T;E), \|\cdot\|_\infty) \hookrightarrow (L_{2,\rho_T}(T;E), \|\cdot\|_{L_{2,\rho_T}(T;E)}). \tag{4.14}$$

As we showed in proposition 3.18, a vector-valued reproducing kernel Hilbert space is a possible candidate for the search space $\mathcal{H}$ if its kernel function $K : T \times T \to \mathcal{L}(E,E)$ can be continuously extended to $\bar{T} \times \bar{T}$.

For the minimization of (A) or (B), respectively, we now restrict ourselves to functions from the bounded ball

$$\mathcal{H}_b := \{f \in \mathcal{H} \mid \|f\|_{\mathcal{H}} \leq b\}$$

for a $b > 0$. Therefore, our setting reads

$$\mathcal{H}_b \subset \mathcal{H} \subset C(T;E) \subset L_{2,\rho_T}(T;E), \tag{4.15}$$

where $\mathcal{H}$ is the search space and $\mathcal{H}_b$ is the search set for the regression problem.

Existence of minimizers

For the existence of a minimizer, it would suffice to prove the compactness of $\mathcal{H}_b$ in $C(T;E)$. In the case where $\mathcal{H}$ is a Hilbert space and E is finite-dimensional, the compactness of $\mathcal{H}_b$ can be proven analogously to the scalar-valued case, see section 2.6 of [23]. To this end, one applies the Arzela-Ascoli theorem for vector-valued functions, see e.g. theorem 17 of chapter 7 in [51], and exploits the fact that weak convergence and strong convergence in $\mathcal{H}$ coincide if E is finite-dimensional. However, in general the requirement of compactness of $\mathcal{H}_b$ in $C(T;E)$ is too restrictive if $\mathcal{H}$ is a reflexive Banach space and E is infinite-dimensional. Therefore, we employ a simplified version of the generalized Weierstrass theorem instead. We begin with the definition of sequential lower semicontinuity.

Definition 4.5 [Sequential lower semicontinuity]
Let $M \subset X$ be a subset of the real Banach space X. A function $F : M \to \mathbb{R}$ is called ***sequentially lower semicontinuous*** *in $u \in M$ if*

$$F(u) \leq \liminf_{n\to\infty} F(u_n)$$

holds for each sequence $(u_n)_{n\in\mathbb{N}} \subset M$ converging to u. We call F sequentially lower semicontinuous on M if it is sequentially lower semicontinuous in every $u \in M$.

Note that continuity obviously implies sequential lower semicontinuity.

Theorem 4.6 [Generalized Weierstrass theorem]
Let $\emptyset \neq M \subset X$ be a subset of the real, reflexive Banach space X. The minimization problem

$$F(u) \to \min_{u\in M}! \tag{4.16}$$

has a solution if M is bounded, closed and convex and if $F : M \to \mathbb{R}$ is convex and sequentially lower semicontinuous.

Proof. See theorem 38.A of [90] in combination with proposition 38.7 and corollary 38.8 of [90]. □

More general versions of this theorem are stated in chapter 38 of [90]. The existence of minimizers for the regression problems (A) and (B) with search set $\mathcal{H}_b$ now follows directly from the generalized Weierstrass theorem.

Corollary 4.7 [Existence of Minimizers of (A) and (B) in $\mathcal{H}_b$]
Let $b > 0$ be arbitrary, let $\mathcal{H}$ be a real, reflexive Banach space which is continuously embedded into $C(T;E)$ with embedding constant $c_{\mathcal{H}}$ and let $\psi(\mathbf{x},\mathbf{y}) = \tilde{\psi}(\mathbf{x}-\mathbf{y})$ be a qualified cost function. Then, the minimization problems

$$\textit{Find } \arg\min_{f\in\mathcal{H}_b} \mathcal{E}(f) \tag{4.17}$$

and

$$\textit{Find} \arg\min_{f \in \mathcal{H}_b} \mathcal{E}_{\mathcal{Z}_n}(f) \tag{4.18}$$

have a minimizer in $\mathcal{H}_b$.

Proof. We apply the generalized Weierstrass theorem 4.6 with $M = \mathcal{H}_b$, $X = \mathcal{H}$ and $F = \mathcal{E}$ or $F = \mathcal{E}_{\mathcal{Z}_n}$, respectively. To this end, we check that all prerequisites are fulfilled. Trivially, $\mathcal{H}_b$ is non-empty, bounded, convex and closed for any $b \geq 0$ and by definition $\mathcal{H}$ is a real, reflexive Banach space.

We now prove the continuity of $\mathcal{E}_{\mathcal{Z}_n}$ on $\mathcal{H}$ and the sequential lower semicontinuity of $\mathcal{E}$ on $\mathcal{H}_b$. Since the embedding $\mathcal{H} \hookrightarrow C(T;E)$ is continuous, the linear point evaluation functionals

$$\delta_{\mathbf{t}} : \mathcal{H} \to E, \quad \delta_{\mathbf{t}}(f) = f(\mathbf{t})$$

are (Lipschitz) continuous on $\mathcal{H}$ for every $\mathbf{t} \in T$. Indeed, we obtain

$$\|\delta_{\mathbf{t}}(f) - \delta_{\mathbf{t}}(g)\|_E \leq \|f - g\|_{L_\infty(T;E)} \leq c_{\mathcal{H}} \|f - g\|_{\mathcal{H}}$$

for all $f, g \in \mathcal{H}$. Together with the continuity of $\tilde{\psi}$, we directly conclude that $\mathcal{E}_{\mathcal{Z}_n}$ is continuous on $\mathcal{H}$ and, therefore, also on $\mathcal{H}_b$ for every $b > 0$. For the sequential lower semicontinuity of $\mathcal{E}$, let $f_n \in \mathcal{H}_b, n \in \mathbb{N}$ converge to $f \in \mathcal{H}_b$ for $n \to \infty$. The functions $G_n : T \times E \to \mathbb{R}$ defined by

$$G_n(\mathbf{t}, \mathbf{x}) := \tilde{\psi}(f_n(\mathbf{t}) - \mathbf{x})$$

are continuous and, therefore, also ρ-measurable for every $n \in \mathbb{N}$. Furthermore, they are non-negative. Hence, by the continuity of the G_n and Fatou's lemma for Lebesgue integrals, see e.g. [3], we obtain

$$\begin{aligned} \mathcal{E}(f) &= \int_{T \times E} \tilde{\psi}(f(\mathbf{t}) - \mathbf{x}) \, \mathrm{d}\rho(\mathbf{t}, \mathbf{x}) = \int_{T \times E} \lim_{n \to \infty} G_n(\mathbf{t}, \mathbf{x}) \, \mathrm{d}\rho(\mathbf{t}, \mathbf{x}) \\ &\leq \liminf_{n \to \infty} \int_{T \times E} G_n(\mathbf{t}, \mathbf{x}) \, \mathrm{d}\rho(\mathbf{t}, \mathbf{x}) = \liminf_{n \to \infty} \mathcal{E}(f_n). \end{aligned}$$

Finally, the convexity of $\mathcal{E}$ and $\mathcal{E}_{\mathcal{Z}_n}$ follows from the linearity of the integral and the point evaluations and from the convexity of $\tilde{\psi}$, i.e. for $\tau \in [0,1]$ and $f_1, f_2 \in \mathcal{H}$, we have

$$\begin{aligned} \mathcal{E}(\tau f_1 + (1-\tau) f_2) &= \int_{T \times E} \tilde{\psi}\left(\tau(f_1(\mathbf{t}) - \mathbf{x}) + (1-\tau)(f_2(\mathbf{t}) - \mathbf{x})\right) \mathrm{d}\rho(\mathbf{t}, \mathbf{x}) \\ &\leq \int_{T \times E} \tau \tilde{\psi}(f_1(\mathbf{t}) - \mathbf{x}) + (1-\tau)\tilde{\psi}(f_2(\mathbf{t}) - \mathbf{x}) \, \mathrm{d}\rho(\mathbf{t}, \mathbf{x}) \\ &= \tau \mathcal{E}(f_1) + (1-\tau)\mathcal{E}(f_2). \end{aligned} \tag{4.19}$$

The same reasoning holds for $\mathcal{E}_{\mathcal{Z}_n}$ and we obtain

$$\begin{aligned}\mathcal{E}_{\mathcal{Z}_n}(\tau f_1 + (1-\tau)f_2) &= \frac{1}{n}\sum_{i=1}^{n}\tilde{\psi}\left(\tau(f_1(\mathbf{t}_i)-\mathbf{x}_i)+(1-\tau)(f_2(\mathbf{t}_i)-\mathbf{x}_i)\right)\\ &\leq \frac{1}{n}\sum_{i=1}^{n}\tau\tilde{\psi}(f_1(\mathbf{t}_i)-\mathbf{x}_i)+(1-\tau)\tilde{\psi}(f_2(\mathbf{t}_i)-\mathbf{x}_i) \qquad (4.20)\\ &= \tau\mathcal{E}_{\mathcal{Z}_n}(f_1)+(1-\tau)\mathcal{E}_{\mathcal{Z}_n}(f_2).\end{aligned}$$

So, both $\mathcal{E}$ and $\mathcal{E}_{\mathcal{Z}_n}$ are convex on $\mathcal{H}$ and thus also on $\mathcal{H}_b$. Therefore, the existence of a minimizer of (4.17) and (4.18) follows from the generalized Weierstrass theorem. □

Uniqueness of minimizers

To derive uniqueness of a minimizer of (4.17) for a qualified cost function $\psi(\mathbf{x},\mathbf{y}) = \tilde{\psi}(\mathbf{x}-\mathbf{y})$, we additionally have to assume that $\tilde{\psi}$ is strictly convex. However, for the minimization of (4.18), this does not suffice since only the point evaluations in $\mathbf{t}_1,\ldots,\mathbf{t}_n$ are relevant to determine the value of $\mathcal{E}_{\mathcal{Z}_n}$.

Proposition 4.8 [Uniqueness of Minimizers of (A) over $\mathcal{H}_b$]
Under the prerequisites of corollary 4.7, the minimizer of (4.17) *is unique if* $\tilde{\psi}$ *is strictly convex on* E*. Furthermore, under these conditions, there exist* $\mathbf{c}_i \in E, i = 1,\ldots,n$ *such that every minimizer* g *of* (4.18) *fulfills* $g(\mathbf{t}_i) = \mathbf{c}_i$*.*

Proof. Due to the strict convexity of $\tilde{\psi}$, the inequality (4.19) becomes strict for every $\tau \in (0,1)$ and all $f_1, f_2 \in \mathcal{H}_b$ for which there exists an $A \subset T$ with $\rho_T(A) > 0$ such that $f_1(\mathbf{t}) \neq f_2(\mathbf{t})$ for all $\mathbf{t} \in A$. Since all functions in $\mathcal{H}_b$ are continuous and ρ_T is non-degenerate, see (4.1), the latter holds for all $f_1, f_2 \in \mathcal{H}_b$ with $f_1 \neq f_2$. Now, assume that $f_1, f_2 \in \mathcal{H}_b$ are two minimizers of (4.17). Then, because of the convexity of $\mathcal{H}_b$, we obtain $\tau f_1 + (1-\tau)f_2 \in \mathcal{H}_b$. Therefore, $f_1 = f_2$ has to hold because otherwise (4.19) would give the strict inequality

$$\mathcal{E}(\tau f_1 + (1-\tau)f_2) < \tau\mathcal{E}(f_1) + (1-\tau)\mathcal{E}(f_2)$$

which would contradict the fact that f_1 and f_2 are minimizers. Thus, the uniqueness of the minimizer of $\mathcal{E}$ over $\mathcal{H}_b$ follows.

Now let us consider $\mathcal{E}_{\mathcal{Z}_n}$. Due to the strict convexity of $\tilde{\psi}$, the inequality (4.20) is strict for every $\tau \in (0,1)$ and all $f_1, f_2 \in \mathcal{H}_b$ for which there exists an $i \in \{1,\ldots,n\}$ such that $f_1(\mathbf{t}_i) \neq f_2(\mathbf{t}_i)$. Thus, by an analogous argumentation as for (4.17), the convexity of $\mathcal{H}_b$ implies the existence of $\mathbf{c}_i \in E$ such that two minimizers $f_1, f_2 \in \mathcal{H}_b$ of (4.18) fulfill $f_1(\mathbf{t}_i) = f_2(\mathbf{t}_i) = \mathbf{c}_i$ for all $i \in \{1,\ldots,n\}$. □

Although (4.18) does not necessarily employ a unique minimizer, the following corollary shows that the minimizer with the smallest $\mathcal{H}$ norm exists and is unique if, additionally, $\mathcal{H}$ is a Hilbert space.

Corollary 4.9 [Uniqueness of minimal norm minimizers of (B) over $\mathcal{H}_b$] *Let the conditions of corollary 4.7 and proposition 4.8 hold. Furthermore, let $\mathcal{H}$ be a Hilbert space. Then, the problem*

$$\textit{Find} \quad \underset{g \in \left\{ \underset{f \in \mathcal{H}_b}{\arg\min}\, \mathcal{E}_{\mathcal{Z}_n}(f) \right\}}{\arg\min} \|g\|_{\mathcal{H}} \tag{4.21}$$

has a unique minimizer.

Proof. To prove the existence of a minimizer of (4.21), we again use the generalized Weierstrass theorem 4.6 with $M = \left\{ \underset{f \in \mathcal{H}_b}{\arg\min}\, \mathcal{E}_{\mathcal{Z}_n}(f) \right\}, X = \mathcal{H}$ and $F(\cdot) = \|\cdot\|_{\mathcal{H}}$. Indeed, the prerequisites of theorem 4.6 are met. To this end, note that $M \subset \mathcal{H}_b$ is bounded in $\mathcal{H}$. Furthermore, it is convex due to the convexity of $\mathcal{H}_b$ and (4.20). Since the embedding (4.14) holds, convergence in $\mathcal{H}$ implies uniform convergence. Using this, together with the fact that

$$M = \{f \in \mathcal{H}_b \mid f(\mathbf{t}_i) = \mathbf{c}_i \; \forall\, i = 1, \ldots, n\}$$

holds for the $\mathbf{c}_i$ from proposition 4.8, the closedness of M follows. By definition, F is continuous in $\mathcal{H}$ and it is also a convex functional. Therefore, the existence of a minimizer of (4.21) follows from theorem 4.6.

Let us now assume that $f_1, f_2 \in M$ are both minimizers of F over M. If $f_1 = 0$, we trivially obtain $f_2 = 0$ since $0 = \|f_1\|_{\mathcal{H}} = \|f_2\|_{\mathcal{H}}$ has to hold. Therefore, we may assume that $f_1, f_2 \neq 0$. Note that this implies the existence of an $i \in \{1, \ldots, n\}$ such that $\mathbf{c}_i \neq 0$ because otherwise $0 \in \mathcal{H}_b$ would be the unique solution to 4.21. Due to the convexity of M and F, $\tau f_1 + (1-\tau) f_2$ is also an element of M and a minimizer of F for every $\tau \in [0, 1]$. Thus, we obtain

$$\|\tau f_1 + (1-\tau) f_2\|_{\mathcal{H}} = \|f_1\|_{\mathcal{H}} = \|f_2\|_{\mathcal{H}} = \tau \|f_1\|_{\mathcal{H}} + (1-\tau)\|f_2\|_{\mathcal{H}}.$$

Since $\mathcal{H}$ is a Hilbert space, it is a strictly convex space, which implies the existence of an $\alpha \in \mathbb{R}$ such that $f_1 = \alpha f_2$, see e.g. [45]. Therefore, we have $\alpha f_2(\mathbf{t}_i) = f_1(\mathbf{t}_i) = f_2(\mathbf{t}_i) = \mathbf{c}_i \neq 0$, which can only be fulfilled if $\alpha = 1$. Thus, $f_1 = f_2$, which proves the uniqueness. □

Choosing $\mathcal{H}_b$ as search set and solving (4.21) corresponds to a specific regularization of the ill-posed problem (B). Note that the minimization of $\mathcal{E}$ over $L_{2,\rho_T}(T; E)$ is already well-posed for the cost function $\psi(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} - \mathbf{y}\|_E^2$ because f_ρ is the unique minimizer according to lemma 4.4. In this case, only the minimization of $\mathcal{E}_{\mathcal{Z}_n}$ has to be regularized.

4.2.2 The error splitting

Definition 4.10 [Notation of minimizers]
For a set $X \subset C(T;E)$ over which minimizers of $\mathcal{E}$ and $\mathcal{E}_{\mathcal{Z}_n}$ exist, e.g. $X = \mathcal{H}_b$, we denote by $f_X \in X$ a minimizer of (A) *over X. Analogously, we denote by $f_{\mathcal{Z}_n,X} \in X$ a minimizer of* (B) *over X.*

As we saw in proposition 4.8, the functions f_X and $f_{\mathcal{Z}_n,X}$ do not have to be unique in general. However, when we use this notation, the function will either be unique or it does not matter which minimizer is taken.

Definition 4.11 [Overall Error]
For a solution $\hat{f}$ of (A) *over $L_{2,\rho_T}(T;E)$, we call*

$$\mathcal{E}\left(f_{\mathcal{Z}_n,\mathcal{H}_b}\right) - \mathcal{E}(\hat{f})$$

*the **overall error** of the regression problem for the search set $\mathcal{H}_b$ and the sample $\mathcal{Z}_n$.*

The overall error measures how much the functional value of $\mathcal{E}$ differs when we take a minimizer of (B) over $\mathcal{H}_b$ instead of considering a direct solution to (A) over $L_{2,\rho_T}(T;E)$. If the minimizer $f_{\mathcal{Z}_n,\mathcal{H}_b}$ is computed by an algorithm, the overall error measures how close the result of the algorithm comes to the true solution in terms of $\mathcal{E}$. Note that this is of course without considering any numerical instabilities, e.g. due to badly conditioned equation systems.

To derive an upper bound for the overall error, we introduce an error splitting into the so-called *bias* and the *sampling error*.

Definition 4.12 [The Bias]
Let $b > 0$ be arbitrary. We call

$$\mathcal{E}(f_{\mathcal{H}_b}) - \mathcal{E}(\hat{f}) \tag{4.22}$$

*the **bias** of the regression problem.*

The bias is a measure for the error that occurs due to the restriction of the search set from the whole space $L_{2,\rho_T}(T;E)$ to $\mathcal{H}_b$, i.e. it indicates how much the minimum of (A) over $\mathcal{H}_b$, cf. (4.17), differs from the minimum of (A) over $L_{2,\rho_T}(T;E)$.

Definition 4.13 [The Sampling Error]
Let $b > 0$ be arbitrary and let $\mathcal{Z}_n \in (T \times E)^n$ be n samples which are drawn according to ρ independently. We call

$$\mathcal{E}(f_{\mathcal{Z}_n,\mathcal{H}_b}) - \mathcal{E}(f_{\mathcal{H}_b}) \tag{4.23}$$

*the **sampling error** of the regression problem.*

The sampling error measures how close a solution of the finite sample regression problem is to a solution of the integral variant of the regression problem in terms of values of the functional $\mathcal{E}$. More precisely, the sampling error is given by the difference between the minimum of (B) over $\mathcal{H}_b$, cf. (4.18), and the minimum of (A) over $\mathcal{H}_b$, cf. (4.17).

The splitting of the overall error into the non-sample-dependent bias and the sampling error for the search set $\mathcal{H}_b$ can be written as

$$\underbrace{\mathcal{E}\left(f_{\mathcal{Z}_n,\mathcal{H}_b}\right) - \mathcal{E}(\hat{f})}_{\text{overall error}} = \underbrace{\mathcal{E}(f_{\mathcal{H}_b}) - \mathcal{E}(\hat{f})}_{\text{bias}} + \underbrace{\mathcal{E}(f_{\mathcal{Z}_n,\mathcal{H}_b}) - \mathcal{E}(f_{\mathcal{H}_b})}_{\text{sampling error}}. \tag{4.24}$$

More details on the actual computation of $f_{\mathcal{Z}_n,\mathcal{H}_b}$ in terms of solving the Lagrangian dual formulation of (4.18) will be discussed in the next chapter.

Summary

In this section, we derived the existence and uniqueness of solutions to the regression problem under certain prerequisites. Furthermore, we defined the overall error and presented a splitting into bias and sampling error. The most important results can be summarized as follows:

- Since we cannot compute a solution $\hat{f}$ to (A) directly, we focus on minimizing (B) over the search set $\mathcal{H}_b$ instead. Here, $\mathcal{H}_b \subset \mathcal{H}$ is a bounded ball in the search space $\mathcal{H}$, for which the embedding (4.14) is continuous. A common example for $\mathcal{H}$ would be a reproducing kernel Hilbert space with continuously extendable kernel function onto $\bar{T} \times \bar{T}$.
- There exist minimizers $f_{\mathcal{H}_b}$ of $\mathcal{E}$ over $\mathcal{H}_b$ and $f_{\mathcal{Z}_n,\mathcal{H}_b}$ of $\mathcal{E}_{\mathcal{Z}_n}$ over $\mathcal{H}_b$ for qualified cost functions such as $\psi(\mathbf{x},\mathbf{y}) = \tilde{\psi}(\mathbf{x}-\mathbf{y}) = \|\mathbf{x}-\mathbf{y}\|_E^2$.
- If $\tilde{\psi}$ is strictly convex, $f_{\mathcal{H}_b}$ is unique. $f_{\mathcal{Z}_n,\mathcal{H}_b}$ is not necessarily unique, but if $\mathcal{H}$ is a Hilbert space, then the $f_{\mathcal{Z}_n,\mathcal{H}_b}$ with the smallest $\mathcal{H}$ norm is unique.
- The overall error $\mathcal{E}\left(f_{\mathcal{Z}_n,\mathcal{H}_b}\right) - \mathcal{E}(\hat{f})$ can be split into the sum of the bias and the sampling error, see (4.24).

4.3 The bias

We now focus on establishing upper bounds on the bias

$$\mathcal{E}(f_{\mathcal{H}_b}) - \mathcal{E}(\hat{f}).$$

First, we consider an infinite-dimensional search space $\mathcal{H}$ which is dense in $L_{2,\rho_T}(T;E)$. In this case, we are able to employ approximation results from real interpolation theory. Based on [23] and our introduction in interpolation theory from section 3.4, we show how the bias can be bounded in this case in subsection 4.3.1. Subsequently, we deal with a finite-dimensional search space V_k instead of $\mathcal{H}$ in subsection 4.3.2. Here, we explain why the known results from subsection 4.3.1 cannot be applied in this case. Afterwards,

we introduce an alternative way to establish an upper bound on the bias by means of Jackson- and Bernstein-type inequalities for the embedding $V_k \hookrightarrow L_{2,\rho_T}(T;E)$.

While the existence and uniqueness results we presented so far hold for arbitrary qualified cost functions, we focus on $\psi(\mathbf{x},\mathbf{y}) = \|\mathbf{x}-\mathbf{y}\|_E^2$ from now on. Therefore, we have $\hat{f} = f_\rho$, see Lemma 4.4.

4.3.1 Infinite-dimensional search spaces

We rely on the results on interpolation spaces from section 3.4. In particular, theorem 3.20 leads to the following bound on the bias.

Corollary 4.14 [Bias Bound for squared norm costs]
Let $b > 0$ and let $\mathcal{H}$ be a reflexive Banach space which fulfills (4.14)*. Let, furthermore, ψ be the squared norm cost function* (4.3) *and let $\sigma \in (0,1)$ be such that $f_\rho \in (L_{2,\rho_T}(T;E),\mathcal{H})_\sigma$. Then, the bias can be bounded by*

$$\mathcal{E}(f_{\mathcal{H}_b}) - \mathcal{E}(f_\rho) \leq \left(\|f\|_\sigma b^{-\sigma}\right)^{\frac{2}{1-\sigma}}. \tag{4.25}$$

Proof. Since $\mathcal{H}$ is reflexive and continuously embedded into $C(T;E)$, the function $f_{\mathcal{H}_b}$ exists due to corollary 4.7. The identity $\mathcal{E}(f_{\mathcal{H}_b}) - \mathcal{E}(f_\rho) = \inf_{f\in\mathcal{H}_b} \|f - f_\rho\|^2_{L_{2,\rho_T}(T;E)}$, see also (4.11), and the application of theorem 3.20 yield the result. □

We observe that the upper bound on the convergence rate of the bias with respect to b is determined by the largest $\sigma \in (0,1)$ such that $f_\rho \in (L_{2,\rho_T}(T;E),\mathcal{H})_\sigma$. For reproducing kernel Hilbert spaces $\mathcal{H}$ of scalar-valued functions for example, the largest such σ is determined by the decay of the eigenvalues of the integral operator which is defined by the kernel function, see e.g. [23] for details.

To illustrate how the above result can be applied, let us consider the example $\mathcal{H} = H^2((0,1);\mathbb{R}^d)$ and $f_\rho \in H^1((0,1);\mathbb{R}^d)$ with $\rho_T = \lambda_T$ being the Lebesgue measure on $T = (0,1)$. Then,

$$f_\rho \in \left(L_2((0,1);\mathbb{R}^d),\mathcal{H}\right)_{\frac{1}{2}},$$

see (3.27). Therefore, according to corollary 4.14, the bias decays like $\mathcal{E}(f_{\mathcal{H}_b}) - \mathcal{E}(f_\rho) = \mathcal{O}\left(b^{-2}\right)$ for $b \to \infty$.

Apart from the bound (4.25) for the search set $\mathcal{H}_b$, there also exist results for more general search sets in L_{2,ρ_T}, see e.g. [8] and [54] for examples in the case $E = \mathbb{R}$. However, we restrict ourselves to the more practical situation where the search set is a bounded ball in $\mathcal{H}$ in this thesis.

4.3.2 Finite-dimensional search spaces

In this section, we point out why we cannot work with the bound (4.25) when dealing with finite-dimensional search spaces. Subsequently, we present a framework which allows to

establish upper bounds on the bias in this case.

Notation for finite-dimensional search spaces

To emphasize that we are dealing with a finite-dimensional search space now, we write V_k for $k \in \mathbb{N}$ instead of $\mathcal{H}$. We denote by $N_k := \dim(V_k) < \infty$ the dimension of V_k. Here, the subscript k refers to the approximation properties of the finite-dimensional space, i.e. the resolution or a comparable quantity. One can think of a chain of finite-dimensional spaces

$$V_0 \subset V_1 \subset V_2 \subset \cdots \tag{4.26}$$

as for Finite-Element discretizations or approximation scales such as wavelet spaces.[3] The larger k gets, the larger N_k becomes. This means that the corresponding spaces V_k are "finer" for larger k. To be precise, we denote by

$$(V_k, \|\cdot\|_{V_k}),\ k \in \mathbb{N}$$

a sequence of finite-dimensional normed subspaces of $C(T;E)$. Although $\|\cdot\|_{V_k}$ could depend on the parameter k, we will neglect this case in this thesis and assume that the norm is fixed for all $k \in \mathbb{N}$. Since all norms on a finite-dimensional vector space are equivalent, we are free to make an adequate choice here. However, the norm serves to regularize the minimization problem in the sense that the search set will be a bounded ball in V_k with respect to this norm. Therefore, it should reflect the regularity assumptions we pose on a minimizer of (4.18). The rest of our notation is completely analogous to the previous sections, i.e. we only substitute $\mathcal{H}$ by V_k. Therefore, all the results from the previous sections, such as statements on existence and uniqueness of solutions of the underlying regression problem for example, are still valid.

To complete the analogy to the notation from the previous sections, we denote the ball of radius b in V_k by

$$V_{k,b} = \{f \in V_k \mid \|f\|_{V_k} \leq b\}.$$

We assume that the embedding $V_k \hookrightarrow C(T;E)$ is continuous with embedding constant

$$c_{V_k} := \|\mathrm{id} : V_k \hookrightarrow C(T;E)\|_{\mathcal{L}(V_k, C(T;E))}.$$

Thus, $V_{k,b}$ fulfills the same prerequisites as $\mathcal{H}_b$ and we have

$$V_{k,b} \subset V_k \subset C(T;E) \subset L_{2,\rho_T}(T;E)$$

in analogy to (4.15). As in the infinite-dimensional case, we call V_k *search space* and $V_{k,b}$ *search set*. In accordance with definition 4.10, we denote the minimizer of (A) over $V_{k,b}$ by $f_{V_{k,b}}$ and we denote a minimizer of (B) over $V_{k,b}$ by $f_{\mathcal{Z}_n, V_{k,b}}$.

[3] Note that we do not necessarily have to follow (4.26), but could also consider non-nested spaces $V_k, k \in \mathbb{N}$.

Interpolation theory and the discretization error

Definition 4.15 [Discretization Error]
Let $b > 0$ be arbitrary. In the case of finite-dimensional search spaces V_k for $k \in \mathbb{N}$, we call the bias

$$\mathcal{E}(f_{V_{k,b}}) - \mathcal{E}(f_\rho) \tag{4.27}$$

*the **discretization error** of the regression problem for resolution k.*

A straightforward way to account for this discretization error would be to apply corollary 4.14. Let us discuss this approach shortly: Since $\dim(V_k) < \infty$ for every $k \in \mathbb{N}$, the search space V_k is not dense in $L_{2,\rho_T}(T;E)$. Therefore, if the solution f_ρ to (A) is not an element of V_k, the $\mathbb{K}$-functional (3.26) for the pair $(L_{2,\rho_T}(T;E), V_k)$ cannot approach zero for $t \to 0$. This means that f_ρ is not an element of the interpolation space $(L_{2,\rho_T}(T;E), V_k)_\sigma$ for any $\sigma > 0$ and we cannot apply corollary 4.14 for a finite-dimensional search space V_k. Therefore, the results on the bias which we presented in subsection 4.3.1 are no longer useful in the case of finite-dimensional search spaces.

The reason why the above-mentioned method fails is quite obvious. We fixed $k \in \mathbb{N}$ and expected to get a convergence of the discretization error to 0 for $b \to \infty$. However, a small discretization error can only be achieved if we choose both b and k large enough.

Jackson- and Bernstein-type inequalities for the discretization error

We now give bounds on the discretization error in the case where certain Jackson and Bernstein inequalities hold. As in (4.24), the overall error is now decomposed into

$$\underbrace{\mathcal{E}(f_{\mathbf{z}_n,V_{k,b}}) - \mathcal{E}(f_\rho)}_{\text{overall error}} = \underbrace{\mathcal{E}(f_{V_{k,b}}) - \mathcal{E}(f_\rho)}_{\text{discretization error}} + \underbrace{\mathcal{E}(f_{\mathbf{z}_n,V_{k,b}}) - \mathcal{E}(f_{V_{k,b}})}_{\text{sampling error}}. \tag{4.28}$$

The basic idea of our approach is to choose the norm bound b such that

$$\inf_{f \in V_{k,b}} \|f - f_\rho\|^2_{L_{2,\rho_T}(T;E)} = \inf_{f \in V_k} \|f - f_\rho\|^2_{L_{2,\rho_T}(T;E)}. \tag{4.29}$$

Then, the discretization error equals the squared $L_{2,\rho_T}(T;E)$ best approximation error in V_k. Now, the norm bound b influences just the sampling error $\mathcal{E}(f_{\mathbf{z}_n,V_{k,b}}) - \mathcal{E}(f_{V_{k,b}})$. As we will see in section 4.4, this term grows if b increases. Therefore, we will choose b as small as possible such that (4.29) is fulfilled. We first define an appropriate L_{2,ρ_T} projector.

Definition 4.16 [Orthogonal projection]
We denote the orthogonal projection onto V_k with respect to the $L_{2,\rho_T}(T;E)$ norm by P_{V_k}, i.e.

$$P_{V_k}(f) := \arg\min_{h \in V_k} \|h - f\|_{L_{2,\rho_T}(T;E)}.$$

Now, we provide a lemma which shows how b has to be chosen with respect to the L_{2,ρ_T} norm of f_ρ.

Lemma 4.17
Let V_k be finite-dimensional search spaces for $k \in \mathbb{N}$ and let $c : \mathbb{N} \to (0, \infty)$ be such that the inverse (Bernstein) inequalities

$$\|f\|_{V_k} \leq c(k) \ \|f\|_{L_{2,\rho_T}(T;E)} \tag{4.30}$$

hold for every $f \in V_k$ and every $k \in \mathbb{N}$. Then, the best approximation $P_{V_k}(f_\rho)$ to f_ρ fulfills

$$\|P_{V_k}(f_\rho)\|_{V_k} \leq c(k) \cdot \|f_\rho\|_{L_{2,\rho_T}(T;E)}.$$

Therefore, we obtain $P_{V_k}(f_\rho) \in V_{k,b}$ if we choose

$$b := c(k) \cdot \|f_\rho\|_{L_{2,\rho_T}(T;E)}. \tag{4.31}$$

Proof. Note that $\langle f_\rho - P_{V_k}(f_\rho), P_{V_k}(f_\rho)\rangle_{L_{2,\rho_T}(T;E)} = 0$ since $\mathrm{Id} - P_{V_k}$ is $L_{2,\rho_T}(T;E)$-orthogonal on V_k. Therefore, it holds

$$\begin{aligned} \|f_\rho\|^2_{L_{2,\rho_T}(T;E)} &= \|f_\rho - P_{V_k}(f_\rho) + P_{V_k}(f_\rho)\|^2_{L_{2,\rho_T}(T;E)} \\ &= \|f_\rho - P_{V_k}(f_\rho)\|^2_{L_{2,\rho_T}(T;E)} + \|P_{V_k}(f_\rho)\|^2_{L_{2,\rho_T}(T;E)} \geq \|P_{V_k}(f_\rho)\|^2_{L_{2,\rho_T}(T;E)} \end{aligned}$$

and we get

$$\|P_{V_k}(f_\rho)\|_{L_{2,\rho_T}(T;E)} \leq \|f_\rho\|_{L_{2,\rho_T}(T;E)}. \tag{4.32}$$

Thus, we obtain the desired result

$$\|P_{V_k}(f_\rho)\|_{V_k} \leq c(k)\|P_{V_k}(f_\rho)\|_{L_{2,\rho_T}(T;E)} \leq c(k)\|f_\rho\|_{L_{2,\rho_T}(T;E)}.$$

using (4.30). □

With the help of lemma 4.17, we can now express the discretization error in terms of the $L_{2,\rho_T}(T;E)$ best approximation error.

Theorem 4.18 [Discretization error for squared norm costs]
Let V_k be a finite-dimensional search space and let $b > 0$ be at least as large as in (4.31). *Let, furthermore, $\psi(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} - \mathbf{y}\|^2_E$ be the squared norm cost function* (4.3). *Then, the discretization error fulfills*

$$\mathcal{E}(f_{V_{k,b}}) - \mathcal{E}(f_\rho) = \inf_{f \in V_k} \|f - f_\rho\|^2_{L_{2,\rho_T}(T;E)}.$$

Proof. Since $P_{V_k}(f_\rho) \in V_{k,b}$ for b greater or equal to (4.31), it obviously minimizes the

$L_{2,\rho_T}(T;E)$ distance to f_ρ among all functions from $V_{k,b}$. Therefore, we have

$$\begin{aligned}\mathcal{E}(f_{V_{k,b}}) - \mathcal{E}(f_\rho) &\overset{\text{lemma 4.4}}{=} \inf_{f\in V_{k,b}} \|f - f_\rho\|^2_{L_{2,\rho_T}(T;E)} \\ &\overset{\text{lemma 4.17}}{=} \|P_{V_k}(f_\rho) - f_\rho\|^2_{L_{2,\rho_T}(T;E)} = \inf_{f\in V_k} \|f - f_\rho\|^2_{L_{2,\rho_T}(T;E)}.\end{aligned}$$

□

We conclude this section with some remarks on theorem 4.18.

- We obtain that the discretization error is equal to the squared $L_{2,\rho_T}(T;E)$ best approximation error to f_ρ for functions in V_k. Therefore, the smoothness of f_ρ has to be exploited by a suitable Jackson-type inequality. We will discuss this in detail when considering several examples in section 4.5.

- The price we paid in theorem 4.18 is the coupling (4.31) between b and $c(k)$. In other words, our results are valid only if b is large enough. As we will see in the next section, the sampling error increases when b becomes larger. Thus, the choice (4.31) is optimal in the sense that b is chosen as small as possible such that theorem 4.18 is still valid.

- Although it might seem unsatisfying at first glance to have a result which only holds under the prerequisite (4.31), we can look at this from a different point of view: Considering a sequence of search spaces V_k for $k \in \mathbb{N}$, the following question arises: Which is the minimal $b > 0$ such that the rate of decay of the discretization error $\mathcal{E}(f_{V_{k,b}}) - \mathcal{E}(f_\rho)$ for $k \to \infty$ is optimal, i.e. when is it equal to the rate of the $L_{2,\rho_T}(T;E)$ best approximation error in V_k? The answer to this question is given by (4.31).

Summary

Let us briefly summarize the results on the bias $\mathcal{E}(f_{\mathcal{H}_b}) - \mathcal{E}(f_\rho)$ or the discretization error $\mathcal{E}(f_{V_{k,b}}) - \mathcal{E}(f_\rho)$, respectively, which we provided in this section.

- If the search space $\mathcal{H}$ is infinite-dimensional and dense in $L_{2,\rho_T}(T;E)$, we can apply corollary 4.14 to obtain a bound on the bias which is governed by the largest $\sigma \in (0,1)$ such that $f_\rho \in (L_{2,\rho_T}(T;E), \mathcal{H})_\sigma$.

- For a finite-dimensional search space V_k, we cannot apply corollary 4.14 anymore and need to take another path to derive a bound on the corresponding discretization error.

- If the parameter $b > 0$ is large enough, theorem 4.18 shows that the discretization error equals the squared $L_{2,\rho_T}(T;E)$ best approximation error in V_k.

4.4 The sampling error

In this section, we consider the sampling error

$$\mathcal{E}(f_{\mathcal{Z}_n,\mathcal{H}_b}) - \mathcal{E}(f_{\mathcal{H}_b})$$

in more detail. Our considerations follow the lines of chapter 3 of [23], where most of the results from this section can be found for the scalar-valued case $E = \mathbb{R}$. Unless explicitly stated otherwise, our results are valid for both finite-dimensional and infinite-dimensional search spaces. Therefore, we will use the general notation $\mathcal{H}$ instead of V_k for the search space again. As we already mentioned in the last section, we focus solely on $\psi(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} - \mathbf{y}\|_E^2$.

Up to now, our search sets did not necessarily need to be compact in $C(T; E)$ in order to prove the existence and uniqueness of minimizers or the bounds on the bias or the discretization error, respectively. To obtain bounds on the sampling error, however, we will consider covering numbers of $\mathcal{H}_b$, which are finite only if $\mathcal{H}_b$ is compact in $C(T; E)$. As we already mentioned in subsection 4.2.1, the vector-valued Arzela-Ascoli theorem, see theorem 17 of chapter 7 in [51], can be used to check for compactness in $C(T; E)$. In particular, if $\mathcal{H}$ is a Hilbert space and $E = \mathbb{R}^d$ for a $d \in \mathbb{N}$, the compactness of $\mathcal{H}_b$ in $C(T; \mathbb{R}^d)$ follows by the same arguments as in the scalar-valued case, see e.g. section 2.6 of [23] for details.

In subsection 4.4.1, we discuss upper bounds on the cost function $\psi(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} - \mathbf{y}\|_E^2$. We proceed by introducing covering numbers and a probabilistic Bernstein inequality in subsection 4.4.2. Subsequently, we provide a lemma on convex search sets, which leads to a theorem on a general upper bound on the sampling error in subsection 4.4.3. We conclude in subsection 4.4.4 with an estimate for covering numbers of finite-dimensional search spaces, which allows for a more detailed analysis of the upper bound derived in 4.4.3.

4.4.1 Bounds on the cost function

First, we need to establish an upper bound on the values which the cost function can attain. To this end, let $b > 0$ be arbitrary. We choose $M_\psi \in (0, \infty)$ such that

$$\psi(f(\mathbf{t}), \mathbf{x}) = \|f(\mathbf{t}) - \mathbf{x}\|_E^2 \leq M_\psi \tag{4.33}$$

holds for ρ-almost every $(\mathbf{t}, \mathbf{x})$ and every $f \in \mathcal{H}_b$. When the condition (4.33) holds, the problem is usually called M**-bounded**, see e.g. [23]. For our specific cost function, we can choose

$$M_\psi = (c_{\mathcal{H}} b + r)^2 \tag{4.34}$$

with embedding constant $c_{\mathcal{H}}$, see (4.13), and r from (4.5) since

$$\|f(\mathbf{t}) - \mathbf{x}\|_E \leq \|f\|_\infty + \|\mathbf{x}\|_E \leq c_{\mathcal{H}}\|f\|_{\mathcal{H}} + r \leq c_{\mathcal{H}} b + r$$

for ρ-almost every $(\mathbf{t}, \mathbf{x})$ and every $f \in \mathcal{H}_b$.

Absolute Bounds on M_ψ

Note that M_ψ from (4.34) depends quadratically on b. This will prove to be a severe limitation when considering optimal convergence rates of the overall error. Therefore, it is often postulated that M_ψ can be chosen independently of b, i.e.

$$M_\psi = (M + r)^2$$

for an absolute constant $0 < M < \infty$. This would be equivalent to taking

$$\{f \in \mathcal{H}_b \mid \|f\|_\infty \leq M\} = \{f \in \mathcal{H} \mid \|f\|_{\mathcal{H}} \leq b,\, \|f\|_\infty \leq M\} \tag{4.35}$$

as a search set. Although it is possible to do so, this significantly changes the corresponding minimization procedure because of the additional constraint. Furthermore, we would need results on the bias/discretization error for this particular search set which cannot be derived that easily. We will neglect this strategy in the following. However, in many situations the functions $f_{\mathcal{Z}_n,\mathcal{H}_b}$ will actually stem from a set of type (4.35) for any $b > 0$ and almost every choice of $\mathcal{Z}_n$. The reason for this is the fact that $\|\mathbf{x}_i\|_E \leq r$ holds almost surely for every $i = 1, \ldots, n$, see (4.5). Thus, $\|f\|_\infty \leq r$ is a meaningful condition for functions f from the search set. Therefore, without explicitly restricting the search set, (4.35) with $M = r$ is often implicitly incorporated as search set in the minimization process of (B) over $\mathcal{H}_b$. In summary, it can be stated that the choice (4.34) might be too pessimistic in many situations. We will also observe this, when we consider numerical experiments in chapter 6.

4.4.2 The probabilistic Bernstein inequality

In this subsection, we introduce the concept of covering numbers. Furthermore, we follow the lines of [23] and provide a vector-valued version of theorem 3.3 specified therein using a probabilistic Bernstein inequality. This Bernstein inequality must not be confused with the inverse inequalities of Bernstein type on which subsection 4.3.2 is based on.

Covering Numbers

The covering number of $\mathcal{H}_b$ with respect to the $L_{\infty,\rho_T}(T; E)$ norm is a crucial ingredient for an upper bound on the sampling error.

Definition 4.19 [Covering numbers and coverings]
*Let A be a compact subset of the Banach space $(Y, \|\cdot\|_Y)$. The **covering number** $\mathcal{N}(A, \varepsilon, Y)$ of A for a radius $\varepsilon > 0$ with respect to Y is defined as the smallest number of balls of radius ε with respect to $\|\cdot\|_Y$ which cover A and whose centers reside in A, i.e.*

$$\mathcal{N}(A, \varepsilon, Y) := \min \left\{ l \in \mathbb{N} \mid \exists\, \mathbf{a}_1, \dots, \mathbf{a}_l \in A : \bigcup_{i=1}^{l} \{\mathbf{y} \in Y \mid \|\mathbf{y} - \mathbf{a}_l\|_Y \leq \varepsilon\} \supseteq A \right\}.$$

*The points $\mathbf{a}_1, \dots, \mathbf{a}_l$ in the above equation are called an ε-**covering** of A with respect to $\|\cdot\|_Y$.*

We could also omit the necessity of compactness in the definition above. However, since the covering number $\mathcal{N}(A, \varepsilon, Y)$ is finite for any $\varepsilon > 0$ if and only if A is compact with respect to Y, it makes sense to already include it in the definition. As we mentioned earlier, one can show that $\mathcal{N}\left(\mathcal{H}_b, \varepsilon, L_{\infty,\rho_T}(T; E)\right)$ is finite if $\mathcal{H}$ is a Hilbert space and $E = \mathbb{R}^d$ for any fixed $d \in \mathbb{N}$ for example.

The probabilistic Bernstein inequality

General bounds on the sampling error can be derived by applying Markov's inequality

$$\mathbb{P}\left[\xi \geq a\right] \leq \frac{\mathbb{E}[\xi]}{a} \quad \forall a > 0,$$

for an appropriate choice of the non-negative random variable ξ, which leads to Bennett's inequality or Hoeffding's inequality for example, see Proposition 3.5 of [23]. The direct application of these inequalities then provides an upper bound on $|\mathcal{E}_{\mathcal{Z}_n}(f) - \mathcal{E}(f)|$ for $f \in \mathcal{H}_b$, from which an upper bound on the sampling error can be derived. However, in our specific situation, where we consider $\psi(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} - \mathbf{y}\|_E^2$, the resulting bound is not optimal in the sense that the convergence rate with respect to the number of sample points n is $n^{-\frac{1}{2}}$ up to logarithms, see e.g. [12] for details. In [23], it is shown for scalar-valued functions that a rate of n^{-1} up to logarithms can be achieved. We now introduce the probabilistic Bernstein inequality which we employ to obtain this rate also in the vector-valued case.

Lemma 4.20
Let S be a set of scalar-valued random variables on $T \times E$. Let there exist $c, B > 0$ such that, for each $\xi \in S$, we have $\mathbb{E}_\rho[\xi] \geq 0$, $\mathbb{E}_\rho[\xi^2] \leq c\mathbb{E}_\rho[\xi]$ and $|\xi - \mathbb{E}_\rho[\xi]| \leq B$ ρ-almost everywhere. Here, the expectation has to be understood with respect to the measure ρ. Then, for every $\eta > 0$ and $0 < \alpha \leq 1$, we obtain

$$\mathbb{P}\left[\sup_{\xi \in S} \frac{\mathbb{E}_\rho[\xi] - \frac{1}{n}\sum_{i=1}^{n} \xi(\mathbf{t}_i, \mathbf{x}_i)}{\sqrt{\mathbb{E}_\rho[\xi] + \eta}} > 4\alpha\sqrt{\eta}\right] \leq \mathcal{N}\left(S, \alpha\eta, L_{\infty,\rho}(T \times E)\right) \exp\left(-\frac{\alpha^2 n \eta}{2c + \frac{2}{3}B}\right). \tag{4.36}$$

Proof. The proof follows from the more general lemma 3.19 of [23]. □

4.4.3 A bound on the sampling error

We now provide a lemma which makes use of the convexity of the search set in order to describe the behavior of $\mathcal{E}(f) - \mathcal{E}(f_{\mathcal{H}_b})$ for an arbitrary $f \in \mathcal{H}_b$. Subsequently, we prove the main result of this section with the help of the probabilistic Bernstein inequality (4.36).

Exploiting the convexity of the search set

Lemma 4.21
Let $b > 0$ and let $f \in \mathcal{H}_b$ be arbitrary. Let furthermore $\psi(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} - \mathbf{y}\|_E^2$. It holds

$$\|f - f_{\mathcal{H}_b}\|^2_{L_{2,\rho_T}(T;E)} \leq \mathcal{E}(f) - \mathcal{E}(f_{\mathcal{H}_b}). \tag{4.37}$$

Proof. The proof is just a vector-valued version of the proof of lemma 3.16 in [23]. To this end, note that $\mathcal{H}_b$ is convex and we have $\tau f + (1-\tau) f_{\mathcal{H}_b} \in \mathcal{H}_b$ for an arbitrary $f \in \mathcal{H}_b$ and $\tau \in [0,1]$. Since $f_{\mathcal{H}_b}$ minimizes $\mathcal{E}$ over $\mathcal{H}_b$, we obtain

$$\begin{aligned}
\|f_{\mathcal{H}_b} - f_\rho\|^2_{L_{2,\rho_T}(T;E)} &\overset{\text{lemma 4.4}}{=} \mathcal{E}(f_{\mathcal{H}_b}) - \mathcal{E}(f_\rho) \\
&\leq \mathcal{E}(\tau f + (1-\tau) f_{\mathcal{H}_b}) - \mathcal{E}(f_\rho) \\
&\overset{\text{lemma 4.4}}{=} \|\tau f + (1-\tau) f_{\mathcal{H}_b} - f_\rho\|^2_{L_{2,\rho_T}(T;E)} \\
&= \|f_{\mathcal{H}_b} - f_\rho\|^2_{L_{2,\rho_T}(T;E)} \\
&+ 2\tau \langle f - f_{\mathcal{H}_b}, f_{\mathcal{H}_b} - f_\rho \rangle_{L_{2,\rho_T}(T;E)} + \tau^2 \|f - f_{\mathcal{H}_b}\|^2_{L_{2,\rho_T}(T;E)}.
\end{aligned}$$

Since τ can be arbitrarily close to 0, we obtain that $K := \langle f - f_{\mathcal{H}_b}, f_{\mathcal{H}_b} - f_\rho \rangle_{L_{2,\rho_T}(T;E)} \geq 0$ has to hold. This leads to

$$\begin{aligned}
\|f - f_\rho\|^2_{L_{2,\rho_T}(T;E)} &= \|f - f_{\mathcal{H}_b}\|^2_{L_{2,\rho_T}(T;E)} + 2K + \|f_{\mathcal{H}_b} - f_\rho\|^2_{L_{2,\rho_T}(T;E)} \\
&\geq \|f - f_{\mathcal{H}_b}\|^2_{L_{2,\rho_T}(T;E)} + \|f_{\mathcal{H}_b} - f_\rho\|^2_{L_{2,\rho_T}(T;E)}.
\end{aligned}$$

Therefore, we have

$$\begin{aligned}
\|f - f_{\mathcal{H}_b}\|^2_{L_{2,\rho_T}(T;E)} &\leq \|f - f_\rho\|^2_{L_{2,\rho_T}(T;E)} - \|f_{\mathcal{H}_b} - f_\rho\|^2_{L_{2,\rho_T}(T;E)} \\
&\overset{\text{lemma 4.4}}{=} \mathcal{E}(f) - \mathcal{E}(f_\rho) - (\mathcal{E}(f_{\mathcal{H}_b}) - \mathcal{E}(f_\rho)) = \mathcal{E}(f) - \mathcal{E}(f_{\mathcal{H}_b}),
\end{aligned}$$

which concludes the proof. □

The proof uses an idea akin to the one in the proof of lemma 4.17. However, there we dealt with a subspace projection, whereas $f_{\mathcal{H}_b}$ can be interpreted as the result of an

application of a convex projection onto $\mathcal{H}_b$.

Note that lemma 4.21 represents a counterpart to lemma 4.3 for the specific choice $f_2 = f_{\mathcal{H}_b}$. However, the square in (4.37) is crucial to obtain the result.

A bound on the sampling error

Using both lemma 4.20 and lemma 4.21, we are able to obtain a probabilistic upper bound on the sampling error $\mathcal{E}(f_{\mathcal{Z}_n,\mathcal{H}_b}) - \mathcal{E}(f_{\mathcal{H}_b})$ for compact search sets.

Theorem 4.22 [Sampling error for squared norm costs]
Let $\mathcal{H}_b$ be a compact subset of $C(T;E)$ and let $\psi(\mathbf{x},\mathbf{y}) = \|\mathbf{x}-\mathbf{y}\|_E^2$. We obtain

$$\mathbb{P}\left[\mathcal{E}(f_{\mathcal{Z}_n,\mathcal{H}_b}) - \mathcal{E}(f_{\mathcal{H}_b}) > \eta\right] \leq \mathcal{N}\left(\mathcal{H}_b, \frac{\eta}{12\sqrt{M_\psi}}, L_{\infty,\rho_T}(T;E)\right) \exp\left(-\frac{n\eta}{300M_\psi}\right) \tag{4.38}$$

for all $\eta > 0$.

Proof. The proof works essentially along the lines of the proof of theorem 3.3 of [23], where $\mathcal{H}$ is a reproducing kernel Hilbert space and only scalar-valued functions are considered. Let

$$S := \left\{\xi(\mathbf{t},\mathbf{x}) := \|f(\mathbf{t}) - \mathbf{x}\|_E^2 - \|f_{\mathcal{H}_b}(\mathbf{t}) - \mathbf{x}\|_E^2 \mid f \in \mathcal{H}_b\right\}.$$

We show that S fulfills the prerequisites of lemma 4.20. To this end, let $\xi \in S$ be arbitrary and let $f \in \mathcal{H}_b$ be the corresponding element of the search set. Then, $\mathbb{E}_\rho[\xi] = \mathcal{E}(f) - \mathcal{E}(f_{\mathcal{H}_b}) \geq 0$ since $f_{\mathcal{H}_b}$ minimizes $\mathcal{E}$ over $\mathcal{H}_b$. Furthermore, we have $-M_\psi \leq \xi(\mathbf{t},\mathbf{x}) \leq M_\psi$ for ρ-almost every $\mathbf{t},\mathbf{x}$ because of (4.33). Therefore, $|\xi(\mathbf{t},\mathbf{x}) - \mathbb{E}_\rho[\xi]| \leq 2M_\psi$ almost everywhere and $B = 2M_\psi$ is a valid choice in lemma 4.20. Finally, note that

$$\begin{aligned} |\xi(\mathbf{t},\mathbf{x})| &= |\langle f(\mathbf{t}) - f_{\mathcal{H}_b}(\mathbf{t}), (f(\mathbf{t}) - \mathbf{x}) + (f_{\mathcal{H}_b}(\mathbf{t}) - \mathbf{x})\rangle_E| \\ &\leq \|f(\mathbf{t}) - f_{\mathcal{H}_b}(\mathbf{t})\|_E \left(\|f(\mathbf{t}) - \mathbf{x}\|_E + \|f_{\mathcal{H}_b}(\mathbf{t}) - \mathbf{x}\|_E\right) \\ &\leq 2\sqrt{M_\psi}\|f(\mathbf{t}) - f_{\mathcal{H}_b}(\mathbf{t})\|_E \end{aligned} \tag{4.39}$$

almost everywhere and, thus,

$$\begin{aligned} \mathbb{E}_\rho\left[\xi^2\right] &\leq 4M_\psi \int_{T\times E} \|f(\mathbf{t}) - f_{\mathcal{H}_b}(\mathbf{t})\|_E^2 \,\mathrm{d}\rho(\mathbf{t},\mathbf{x}) \\ &\overset{(4.2)}{=} 4M_\psi \int_T \|f(\mathbf{t}) - f_{\mathcal{H}_b}(\mathbf{t})\|_E^2 \,\mathrm{d}\rho_T(\mathbf{t}) \\ &\overset{\text{lemma 4.21}}{\leq} 4M_\psi \left(\mathcal{E}(f) - \mathcal{E}(f_{\mathcal{H}_b})\right) = 4M_\psi \mathbb{E}_\rho[\xi]. \end{aligned}$$

Therefore, all prerequisites of lemma 4.20 are fulfilled and we take $c = 4M_\psi$ and $B = 2M_\psi$

there to obtain

$$\begin{aligned}
&\mathbb{P}\left[\sup_{f\in\mathcal{H}_b}\frac{\mathcal{E}(f)-\mathcal{E}(f_{\mathcal{H}_b})-\mathcal{E}_{\mathcal{Z}_n}(f)+\mathcal{E}_{\mathcal{Z}_n}(f_{\mathcal{H}_b})}{\sqrt{\mathcal{E}(f)-\mathcal{E}(f_{\mathcal{H}_b})+\eta}}>4\alpha\sqrt{\eta}\right] \qquad (4.40)\\
\leq\ &\mathcal{N}\left(S,\alpha\eta,L_{\infty,\rho}(T\times E)\right)\exp\left(-\frac{\alpha^2 n\eta}{\frac{28}{3}M_\psi}\right)\\
\leq\ &\mathcal{N}\left(\mathcal{H}_b,\frac{\alpha\eta}{2\sqrt{M_\psi}},L_{\infty,\rho_T}(T;E)\right)\exp\left(-\frac{\alpha^2 n\eta}{\frac{28}{3}M_\psi}\right),
\end{aligned}$$

where the last inequality follows from the fact that each ε-covering of $\mathcal{H}_b$ with respect to the $L_{\infty,\rho_T}(T;E)$ norm corresponds to a $2\sqrt{M_\psi}\varepsilon$-covering of S with respect to the $L_{\infty,\rho}(T\times E)$ norm for an arbitrary $\varepsilon>0$. Indeed, let $f_1,f_2\in\mathcal{H}_b$ be arbitrary and let $\xi_1,\xi_2\in S$ be the corresponding functions in S. Then, similarly as in (4.39), we obtain

$$\begin{aligned}
\|\xi_1-\xi_2\|_{L_{\infty,\rho}(T\times E)} &= \operatorname*{ess\,sup}_{\mathbf{t}\in T,\mathbf{x}\in E}\left|\langle f_1(\mathbf{t})-f_2(\mathbf{t}),(f_1(\mathbf{t})-\mathbf{x})+(f_2(\mathbf{t})-\mathbf{x})\rangle_E\right|\\
&\leq \operatorname*{ess\,sup}_{\mathbf{t}\in T,\mathbf{x}\in E}\|f_1(\mathbf{t})-f_2(\mathbf{t})\|_E\left(\|f_1(\mathbf{t})-\mathbf{x}\|_E+\|f_2(\mathbf{t})-\mathbf{x}\|_E\right)\\
&\leq 2\sqrt{M_\psi}\operatorname*{ess\,sup}_{\mathbf{t}\in T}\|f_1(\mathbf{t})-f_2(\mathbf{t})\|_E=2\sqrt{M_\psi}\|f_1-f_2\|_{L_{\infty,\rho_T}(T;E)}.
\end{aligned}$$

Now, choosing $f=f_{\mathcal{Z}_n,\mathcal{H}_b}$ and $\alpha=\frac{\sqrt{2}}{8}$ in (4.40) and setting $E:=\mathcal{E}(f_{\mathcal{Z}_n,\mathcal{H}_b})-\mathcal{E}(f_{\mathcal{H}_b})$, we observe

$$\begin{aligned}
&\mathbb{P}\left[E>\sqrt{\frac{\eta}{2}}\sqrt{E+\eta}+\mathcal{E}_{\mathcal{Z}_n}(f_{\mathcal{Z}_n,\mathcal{H}_b})-\mathcal{E}_{\mathcal{Z}_n}(f_{\mathcal{H}_b})\right] \qquad (4.41)\\
\leq\ &\mathcal{N}\left(\mathcal{H}_b,\frac{\sqrt{2}\eta}{16\sqrt{M_\psi}},L_{\infty,\rho_T}(T;E)\right)\exp\left(-\frac{3n\eta}{896M_\psi}\right)\\
\leq\ &\mathcal{N}\left(\mathcal{H}_b,\frac{\eta}{12\sqrt{M_\psi}},L_{\infty,\rho_T}(T;E)\right)\exp\left(-\frac{n\eta}{300M_\psi}\right)=:A
\end{aligned}$$

since reducing the radius increases the covering number. Note that $\mathcal{E}_{\mathcal{Z}_n}(f_{\mathcal{Z}_n,\mathcal{H}_b})-\mathcal{E}_{\mathcal{Z}_n}(f_{\mathcal{H}_b})\leq 0$ because $f_{\mathcal{Z}_n,\mathcal{H}_b}$ minimizes $\mathcal{E}_{\mathcal{Z}_n}$ over $\mathcal{H}_b$. Thus, by omitting this term in (4.41) the probability gets smaller. Therefore, $\mathbb{P}\left[E>\sqrt{\frac{\eta}{2}}\sqrt{E+\eta}\right]\leq A$. Finally, note that

$$\begin{aligned}
E>\sqrt{\frac{\eta}{2}}\sqrt{E+\eta}\ &\Leftrightarrow E^2>\frac{\eta}{2}E+\frac{\eta^2}{2}\ \Leftrightarrow\ \left(E-\frac{\eta}{4}\right)^2>\frac{9}{16}\eta^2\\
&\Leftrightarrow E-\frac{1}{4}\eta>\frac{3}{4}\eta\ \text{ or }\ E-\frac{1}{4}\eta<-\frac{3}{4}\eta\ \Leftrightarrow\ E>\eta\ \text{ or }\ E<-\frac{1}{2}\eta.
\end{aligned}$$

Since $E = \mathcal{E}(f_{\mathcal{Z}_n,\mathcal{H}_b}) - \mathcal{E}(f_{\mathcal{H}_b}) \geq 0$ by definition, we obtain $\mathbb{P}[E > \eta] \leq A$, which completes the proof. □

Note that a result based on the covering number with respect to the $L_{2,\rho_T}(T;E)$ norm instead of the $L_{\infty,\rho_T}(T;E)$ norm would be more natural since this reflects the norm in which the overall error is measured, see lemma 4.4. However, it has been shown in [54] that such a result cannot hold in full generality. Note, furthermore, that lemma 4.21 and, therefore, also theorem 4.22 implicitly exploit the convexity of $\mathcal{H}_b$. Nevertheless, similar results also hold in the case of non-convex search sets under the premise that the bias is small enough, see theorem 3.2 of [79].

4.4.4 The sampling error for finite-dimensional search spaces

For finite-dimensional search spaces $\mathcal{H} = V_k$ with $k \in \mathbb{N}$, we can estimate the covering number on the right-hand side of (4.38) directly, see also [12]. As mentioned before, we will use $N_k := \dim(V_k)$ to denote the degrees of freedom of V_k.

Lemma 4.23
Let $\eta, b > 0$ and let $M_\psi > 0$ fulfill (4.33). If $b \geq c_1 > 0$ and $\eta \leq c_2 < \infty$, there exists a constant $c_{\mathcal{N}}$ such that

$$\mathcal{N}\left(V_{k,b}, \frac{\eta}{12\sqrt{M_\psi}}, L_\infty(T;E)\right) \leq \left(\frac{c_{\mathcal{N}} c_{V_k} \sqrt{M_\psi} b}{\eta}\right)^{N_k}$$

for any finite-dimensional search space V_k.

Proof. Since V_k is continuously embedded into $C(T;E)$, we know that $\|\cdot\|_{L_{\infty,\rho_T}(T;E)} \leq c_{V_k}\|\cdot\|_{V_k}$ holds for all functions in V_k. Therefore, an ε-covering of V_k with respect to $\|\cdot\|_{V_k}$ is a $c_{V_k}\varepsilon$-covering of V_k with respect to the $L_\infty(T;E)$ norm for an arbitrary $\varepsilon > 0$. Thus, we have

$$\mathcal{N}\left(V_{k,b}, \frac{\eta}{12\sqrt{M_\psi}}, L_\infty(T;E)\right) \leq \mathcal{N}\left(V_{k,b}, \frac{\eta}{12 c_{V_k}\sqrt{M_\psi}}, V_k\right) \leq \left(\frac{24 c_{V_k}\sqrt{M_\psi} b}{\eta} + 1\right)^{N_k}.$$

The last inequality is provided in theorem 5.3 of [23] and holds for any finite-dimensional Banach space V_k. Since $c_{V_k} \geq 1$ by definition and since M_ψ and $\frac{b}{\eta}$ are bounded from below, there exists a $c_{\mathcal{N}} > 0$ such that

$$\left(\frac{24 c_{V_k}\sqrt{M_\psi} b}{\eta} + 1\right)^{N_k} \leq \left(\frac{c_{\mathcal{N}} c_{V_k}\sqrt{M_\psi} b}{\eta}\right)^{N_k},$$

which completes the proof. □

Note that the prerequisites of lemma 4.23 are no restriction for our analysis since we are interested in the case $b \to \infty$ and $\eta \to 0$ anyhow. With the help of this result on the covering number, we obtain the following estimate on the sampling error.

Theorem 4.24 [The sampling error for finite-dimensional search spaces] *Let the prerequisites of lemma 4.23 be fulfilled, let $(V_k)_{k=1}^{\infty}$ be a sequence of finite-dimensional search spaces and let $\psi(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} - \mathbf{y}\|_E^2$. Let furthermore $0 < \delta < 1$ be fixed. Then, we have*

$$\mathcal{E}(f_{\mathcal{Z}_n, V_{k,b}}) - \mathcal{E}(f_{V_{k,b}}) \leq \frac{300 M_\psi N_k}{n} \max\left(1, \log\left(\frac{c_{\mathcal{N}} c_{V_k} bn}{300\,\delta \sqrt{M_\psi} N_k}\right)\right) \tag{4.42}$$

with probability at least $1 - \delta$.

Proof. By equating

$$\delta = \left(\frac{c_{\mathcal{N}} c_{V_k} \sqrt{M_\psi} b}{\eta}\right)^{N_k} \exp\left(-\frac{n\eta}{300 M_\psi}\right), \tag{4.43}$$

we know from lemma 4.23 and theorem 4.22 that

$$\mathbb{P}\left[\mathcal{E}(f_{\mathcal{Z}_n, V_{k,b}}) - \mathcal{E}(f_{V_{k,b}}) \leq \eta\right] \geq 1 - \delta.$$

Therefore, to complete the proof, it suffices to show that η is smaller or equal to the right-hand side of (4.42) if (4.43) holds. To this end, note that we obtain

$$\begin{aligned} \exp\left(\frac{n\eta}{300 M_\psi}\right) \eta^{N_k} &= \left(c_{\mathcal{N}} c_{V_k} \sqrt{M_\psi}\, b\, \delta^{-\frac{1}{N_k}}\right)^{N_k} \\ \Leftrightarrow \exp(\alpha\eta)\, \eta &= \beta \end{aligned} \tag{4.44}$$

with $\alpha := \frac{n}{300 M_\psi N_k}$ and $\beta := c_{\mathcal{N}} c_{V_k} \sqrt{M_\psi}\, b\, \delta^{-\frac{1}{N_k}}$ by reformulating (4.43). Next, we multiply both sides of the equation by α and apply the monotone increasing Lambert W-function $W : [0, \infty) \to [0, \infty)$ defined by

$$W(t \exp(t)) := t$$

on both sides of (4.44). Thus, we have[4]

$$\begin{aligned} & \alpha\eta = W(\alpha\beta) \\ \Leftrightarrow \quad & \eta = \frac{1}{\alpha} W(\alpha\beta) \leq \frac{1}{\alpha} \max(1, \log(\alpha\beta)) \end{aligned} \tag{4.45}$$

[4]Note that this also proves that the assignment (4.43) is valid because it is equivalent to (4.45), i.e. for each δ there exists exactly one η and vice versa.

since $W(s) \leq \log(s)$ for all $s \geq \exp(1)$. Writing out (4.45), we obtain that

$$\mathcal{E}(f_{\mathcal{Z}_n,V_{k,b}}) - \mathcal{E}(f_{V_{k,b}}) \leq \frac{300 M_\psi N_k}{n} \max\left(1, \log\left(\frac{n c_{\mathcal{N}} c_{V_k} \sqrt{M_\psi}\, b\, \delta^{-\frac{1}{N_k}}}{300 M_\psi N_k}\right)\right)$$

holds with probability at least $1-\delta$. To conclude the proof, note that $\delta^{-\frac{1}{N_k}} \leq \delta^{-1}$ for all N_k since $0 < \delta < 1$ and $N_k = \dim(V_k) \geq 1$. □

Summary

In this section, we presented probabilistic upper bounds on the sampling error $\mathcal{E}(f_{\mathcal{Z}_n,\mathcal{H}_b}) - \mathcal{E}(f_{\mathcal{H}_b})$. We conclude with a summary of the obtained results.

- It is necessary to ensure that there exists $M_\psi > 0$ which fulfills the so-called M-boundedness condition (4.33) in order to work with covering number results on compact search sets $\mathcal{H}_b$. While absolute bounds which are independent of b, i.e. $M_\psi \simeq 1$, can usually not be obtained for every function from $\mathcal{H}_b$, it can often be observed that these absolute bounds are valid for the minimizers of (B) over $\mathcal{H}_b$ for arbitrary b and $\mathcal{Z}_n$. This often leads to a favorable behavior of the convergence rates of the sampling error in numerical experiments.
- With the help of the probabilistic Bernstein inequality (4.36) and the convexity of the search set, we derived the bound (4.38) for the sampling error.
- For the special case of finite-dimensional search sets V_k, we obtain the upper bound (4.42) with fixed confidence $1-\delta$. It depends on the sample size n, the cost function barrier M_ψ, the search set radius b and the dimension N_k of V_k.

4.5 Examples for constrained regression

We now consider four examples to demonstrate how the above results can be used to obtain error bounds for constrained regression in specific settings. Since we focus on regression over finite-dimensional search sets in this thesis, we will not discuss the infinite-dimensional case. However, the recipe to deal with this case should be clear by now: One needs an interpolation result of type $f_\rho \in (L_{2,\rho_T}(T;E), \mathcal{H})_\sigma$ to apply corollary 4.14 and obtain an estimate on the bias as we did in subsection 4.3.1. Subsequently, a bound on the covering number for balls in the infinite-dimensional search set is needed. To this end, we refer to [28, 86]. For a thorough analysis of examples for regression over infinite-dimensional search spaces, we refer the interested reader to [23], which is devoted solely to this topic.

In this section, we first deal with piecewise linear splines on full grids and sparse grids as introduced in subsection 3.5.1. Furthermore, we consider Fourier polynomials

on full grids and hyperbolic crosses, which we reviewed in subsection 3.5.2. Grid-based algorithms are motivated by the fact that they can get rid of the cubic computational costs with respect to the amount n of data points, which data-based approaches usually suffer from, see e.g. chapter 10 of [74]. Therefore, grid-based search sets are a good alternative to so-called kernel methods, especially if the dimension m of the problem is low and the number of data n is quite large. We will have a more detailed look on this issue in subsection 5.1.2. However, due to the curse of dimensionality, full grid methods are no longer feasible if $m > 3$. Therefore, sparse grids and hyperbolic crosses need to be employed for moderate-dimensional cases, i.e. up to $m = 10$. In recent years, regression methods based on these spaces have been successfully applied, see e.g. [11, 68].

For our examples, we stick to $E = \mathbb{R}^d$ in the spline case and $E = \mathbb{C}^d$ in the Fourier case, respectively, with $d \in \mathbb{N}$. Furthermore, we use the squared norm cost function $\psi(\mathbf{x}, \mathbf{y}) := \|\mathbf{x} - \mathbf{y}\|_E^2$. We consider scales of finite-dimensional search spaces $(V_k)_{k=0}^{\infty}$ and the corresponding balls $V_{k,b}$. As we already mentioned in subsection 4.4.1, it makes sense to consider both cases

$$M_\psi = (M + r)^2 \simeq 1 \quad \text{and} \quad M_\psi = (c_{V_k} b + r)^2 \simeq c_{V_k}^2 b^2,$$

for the cost function bound M_ψ, see also (4.33). Here, the constants[5] implied by $\simeq$ do not depend on k.

The regularization norm $\|\cdot\|_{V_k}$ is deliberately chosen to be the H^1 norm in the full grid case and the H^1_{mix} norm for sparse grids and hyperbolic crosses because these norms fit both the choice of our basis functions, see also [11], and the norm equivalences we provided in section 3.5. For Fourier polynomials on hyperbolic crosses, one could also consider higher degree Sobolev norms such as the $H^{\hat{s}}(T; \mathbb{C}^d)$ and $H^{\hat{s}}_{\text{mix}}(T; \mathbb{C}^d)$ norms with $\hat{s} \geq 2$. However, we focus on $\hat{s} = 1$ in our examples.

We explain our methodology in more detail for the first example and give a more brief explanation for the subsequent ones. Although we do not discuss the constants in front of the rates which we derive for the overall error, we show at least the explicit dependence of the error on the confidence level $1 - \delta$, which we keep fixed.

The general procedure is the same for all examples in the subsections 4.5.1 - 4.5.4:

- We determine the scaling of the embedding constant c_{V_k}.
- With the help of an inverse inequality of Bernstein type, we determine the growth of the regularization parameter b.
- An application of theorem 4.18 to estimate the discretization error and theorem 4.24 to estimate the sampling error provides an upper bound on the overall error.

[5] Note that we implicitly assume that b is bounded away from zero, i.e. there exists a $c > 0$ such that $b > c$. Otherwise $(c_{V_k} b + r)^2 \simeq c_{V_k}^2 b^2$ does not hold for b arbitrarily close to 0. However, this is no restriction at all since we are interested in the case $b \to \infty$ anyhow.

- We determine necessary conditions as well as sufficient conditions to obtain convergence of the overall error.

- We discuss the optimal coupling between the number of sample points n and the degrees of freedom N_k which balances the discretization error and the sampling error and provide the corresponding convergence rate with respect to n.

In order to summarize the results, we give a comparison and an overview on the considered settings in subsection 4.5.5. In the final subsection 4.5.6, we relate our findings to other research results in this direction.

4.5.1 Regression with piecewise linear basis functions on full grids

First, we consider multivariate regression on full grids with piecewise linear basis functions, see also subsection 3.5.1. To this end, let $T = (0,1)^m$ and let $\rho_T = \lambda_T$ be the Lebesgue measure. As already mentioned in chapter 3, we use the notation $L_2(T;E)$ for $L_{2,\rho_T}(T;E)$ in this case. We assume that $f_\rho \in H^s(T;\mathbb{R}^d)$ for a $0 < s \leq 2$. Let $V_k = \mathcal{V}_k^{\text{full},d}$ be the prewavelet space on a full grid of level k, see (3.33). Note that a properly adjusted ansatz space might be more appropriate if f_ρ stems from a Sobolev space with a higher degree of smoothness. One can e.g. employ grids which are based on higher order splines in such a case, see [13].

As mentioned above, we choose $\|\cdot\|_{V_k} = \|\cdot\|_{H^1(T;\mathbb{R}^d)}$. If $m = 1$, proposition 3.18 ensures that

$$\exists\, c > 0 : \|f\|_\infty \leq c\|f\|_{H^1(T;\mathbb{R}^d)}.$$

Thus, the embedding constant $c_{V_k} = c \simeq 1$ of

$$(V_k, \|\cdot\|_{V_k}) \hookrightarrow \left(C(T;\mathbb{R}^d), \|\cdot\|_\infty\right)$$

is bounded from above independently from k. If $m \geq 2$, however, we know from subsection 3.3.2 that $H^1(T;\mathbb{R}^d)$ is not a reproducing kernel Hilbert space anymore and we cannot assume that c_{V_k} is independent from k. To obtain a bound on c_{V_k} in this case, let $f \in V_k$ be given by

$$f = \sum_{j=1}^{d} \sum_{|\mathbf{l}|_{\ell_\infty} \leq k} \sum_{\mathbf{i} \in \mathbf{I}_\mathbf{l}} \alpha_{\mathbf{l},\mathbf{i},j} \gamma_{\mathbf{l},\mathbf{i}} \mathbf{e}_j.$$

Let, furthermore, $\vec{\alpha} \in \mathbb{R}^{N_k}$ be the vector of all coefficients $\alpha_{\mathbf{l},\mathbf{i},j}$. Note that

$$\sup_{\mathbf{t} \in T} |\gamma_{\mathbf{l},\mathbf{i}}(\mathbf{t})| \leq \left(\frac{6}{5}\right)^m 2^{\frac{km}{2}}$$

holds for each prewavelet basis function, which can easily be seen from their definition

(3.29). Then, with $N_k = d(2^k+1)^m \simeq 2^{km}$, we obtain

$$\begin{aligned}\|f\|_\infty &\leq \sum_{j=1}^{d} \sum_{|\mathbf{l}|_{\ell_\infty}\leq k} \sum_{\mathbf{i}\in\mathbf{I}_\mathbf{l}} |\alpha_{\mathbf{l},\mathbf{i},j}| \|\gamma_{\mathbf{l},\mathbf{i}}\mathbf{e}_j\|_\infty \leq \left(\frac{6}{5}\right)^m 2^{\frac{km}{2}} \cdot \|\vec{\alpha}\|_{\ell_1} \\ &\leq \left(\frac{6}{5}\right)^m 2^{\frac{km}{2}} \sqrt{N_k} \|\vec{\alpha}\|_{\ell_2} \lesssim 2^{km} \|\vec{\alpha}\|_{\ell_2} \lesssim 2^{km} \|f\|_{L_2(T;\mathbb{R}^d)} \leq 2^{km} \|f\|_{H^1(T;\mathbb{R}^d)},\end{aligned}$$

where we used the Riesz stability of the prewavelet basis functions, i.e. (3.37) with $s = 0$, to obtain the relation $\|\vec{\alpha}\|_{\ell_2} \lesssim \|f\|_{L_2(T;\mathbb{R}^d)}$, see also [41, 42]. Recall that $x(k) \lesssim y(k)$ indicates that there exists a constant $c > 0$, which is independent of k, such that $x(k) \leq cy(k)$ for all $k \in \mathbb{N}$. Therefore, we can choose $c_{V_k} \lesssim 2^{km}$ if $m \geq 2$. Note, however, that this might be a very crude estimate as we essentially used only a Nikolskii-type inequality between the N_k-dimensional ℓ_1 and ℓ_2 spaces, which did not make use of the smoothness implied by the H^1 norm.

Because of our differentiation between $m = 1$ and $m > 1$ and also between $M_\psi \simeq 1$ and $M_\psi \simeq c_{V_k}^2 b^2$, we would have to make a four-fold case analysis for our further considerations. For the ease of notation and for the sake of readability, we omit the case $m = 1$ in the following. Note, however, that this case coincides with $m = 1$ for sparse grids, which we will deal with in our next example.

The overall error

To estimate the discretization error, we want to apply theorem 4.18 and need an appropriate inverse inequality. To this end, we use the Bernstein-type inequality (3.39), which states that there exists an m-dependent constant $\tilde{c} > 0$ such that

$$\|f\|_{H^1(T;\mathbb{R}^d)} \leq \tilde{c} 2^k \|f\|_{L_2(T;\mathbb{R}^d)} \quad \forall\, f \in V_k.$$

Thus, $b := \tilde{c} 2^k r \geq \tilde{c} 2^k \|f_\rho\|_{L_2(T;\mathbb{R}^d)}$ with r from (4.5) is a valid choice in theorem 4.18, which leads to

$$\mathcal{E}(f_{V_{k,b}}) - \mathcal{E}(f_\rho) = \inf_{f\in V_k} \|f - f_\rho\|_{L_2(T;\mathbb{R}^d)}^2 \overset{(3.41)}{=} \mathcal{O}\left(2^{-2sk}\right)$$

for full grids, where the $\mathcal{O}$-term has to be understood for $k \to \infty$ and the implicit constant depends (exponentially) on s and m and (linearly) on d. Together with the sampling error bound from theorem 4.24, we obtain the overall error

$$\mathcal{E}(f_{\mathcal{Z}_n,V_{k,b}}) - \mathcal{E}(f_\rho) \lesssim 2^{-2sk} + \frac{M_\psi N_k}{n} \max\left(1, \log\left(\frac{c_{V_k} bn}{\delta\sqrt{M_\psi N_k}}\right)\right),$$

with confidence $1-\delta$, where we already used the fact that $c_\mathcal{N}$ from (4.42) is independent of k and n. Since we only analyze the case $m \geq 2$ here, we use $c_{V_k} \simeq 2^{km}$, $b \simeq 2^k$ and

$N_k \simeq 2^{km}$ to obtain

$$\mathcal{E}(f_{\mathcal{Z}_n,V_{k,b}}) - \mathcal{E}(f_\rho) \lesssim 2^{-2sk} + \frac{2^{km}}{n} \max\left(1, \log\left(\frac{2^k n}{\delta}\right)\right)$$

if $M_\psi = (M+r)^2 \simeq 1$ and

$$\mathcal{E}(f_{\mathcal{Z}_n,V_{k,b}}) - \mathcal{E}(f_\rho) \lesssim 2^{-2sk} + \frac{2^{2k(\frac{3}{2}m+1)}}{n} \max\left(1, \log\left(\frac{n}{\delta 2^{km}}\right)\right)$$

if $M_\psi = (c_{V_k} b + r)^2 \simeq 2^{2k(m+1)}$. Here, the constants which are implied by $\simeq$ and $\lesssim$ depend on m, d and s.

The convergent case

Looking at the previous results in more detail, we see that the first summand in both rates converges to 0 for $k \to \infty$. For the second summand, however, we have to impose an additional condition on the coupling between n and k to obtain convergence. To this end, note that $n > 2^{km}$ is necessary to obtain convergence for both choices of M_ψ. This implies that the max term evaluates to the second argument, i.e. if $n > 2^{km}$, we get

$$\mathcal{E}(f_{\mathcal{Z}_n,V_{k,b}}) - \mathcal{E}(f_\rho) \lesssim 2^{-2sk} + \frac{2^{km}}{n} \log\left(\frac{2^k n}{\delta}\right) \lesssim 2^{-2sk} + \frac{2^{km}}{n} \log\left(\frac{n}{\delta}\right) \tag{4.46}$$

for $M_\psi \simeq 1$, where we used $\log(2^k n) \leq \log(n^2) = 2\log(n)$ for the last inequality, and

$$\mathcal{E}(f_{\mathcal{Z}_n,V_{k,b}}) - \mathcal{E}(f_\rho) \lesssim 2^{-2sk} + \frac{2^{2k(\frac{3}{2}m+1)}}{n} \log\left(\frac{n}{\delta 2^{km}}\right) \tag{4.47}$$

for $M_\psi \simeq 2^{2k(m+1)}$. Note that the sample size n implicitly suffers from the curse of dimensionality due to the necessary condition $n > 2^{km}$.

We now consider sufficient conditions for the convergence of (4.46) and (4.47). As explained in chapter 2, we write $x(k) \ll y(n)$ to state that $k = k(n)$ is coupled to n in such a way that $k(n) \to \infty$ for $n \to \infty$ and $x(k(n)) = o(y(n))$, where o denotes the little-o Landau symbol. For fixed confidence $1-\delta$, we observe that

$$N_k \ll \frac{n}{\log(n)}$$

with $N_k \simeq 2^{km}$ is a sufficient condition to obtain convergence of the right-hand side of (4.46) to 0. When considering (4.47), the condition for convergence to 0 becomes more severe, i.e. here it suffices to ensure

$$2^{2k} (N_k)^3 \ll \frac{n}{\log(n)}.$$

Balancing the error terms

To find the optimal scaling between the grid level k and the sample size n, we (approximately) balance the discretization error and the sampling error by equating the summands in the error estimates from the last paragraph. With a fixed confidence $1-\delta$, we obtain

$$2^{-2sk} \simeq \frac{2^{km}}{n} \log\left(\frac{n}{\delta}\right)$$

for (4.46). This can be reformulated as $n \simeq 2^{(2s+m)k} \log\left(\frac{n}{\delta}\right)$, which is essentially

$$n \simeq 2^{(2s+m)k} \simeq 2^{2sk} N_k \simeq N_k^{\frac{2s+m}{m}} \tag{4.48}$$

up to logarithms in n. Substituting $2^k \simeq n^{\frac{1}{2s+m}}$, which follows from (4.48), into (4.46), we obtain

$$\mathcal{E}(f_{\mathcal{Z}_n,V_{k,b}}) - \mathcal{E}(f_\rho) \lesssim n^{-\frac{2s}{2s+m}} + \frac{n^{\frac{m}{2s+m}}}{n} \log\left(\frac{n}{\delta}\right) = \mathcal{O}\left(n^{-\frac{2s}{2s+m}} \log(n)\right) \tag{4.49}$$

for $n \to \infty$. Therefore, for the coupling (4.48), our convergence rate estimate for multivariate regression on full grids with piecewise linear basis functions in the case of M-boundedness with $M_\psi \simeq 1$ is (4.49).

In the case $M_\psi \simeq (c_{V_k} b + r)^2$, in which we have (4.47), we obtain the balanced scaling

$$2^{-2sk} \simeq \frac{2^{2k(\frac{3}{2}m+1)}}{n} \log\left(\frac{n}{\delta 2^{km}}\right) \;\Leftrightarrow\; n \simeq 2^{(2s+3m+2)k} \log\left(\frac{n}{\delta 2^{km}}\right).$$

Substituting $2^k = n^{\frac{1}{2s+3m+2}}$ into (4.47), we finally get

$$\begin{aligned}\mathcal{E}(f_{\mathcal{Z}_n,V_{k,b}}) - \mathcal{E}(f_\rho) &\lesssim n^{-\frac{2s}{2s+3m+2}} \left(1 + \frac{2(s+m+1)}{2s+3m+2} \log(n) - \log(\delta)\right) \\ &= \mathcal{O}\left(n^{-\frac{2s}{2s+3m+2}} \log(n)\right).\end{aligned} \tag{4.50}$$

4.5.2 Regression with piecewise linear basis functions on sparse grids

We now have a look at multivariate regression on sparse grids with the piecewise linear prewavelet basis functions. We will see that, in contrast to full grids, the necessary number of samples to obtain convergence does not suffer from the curse of dimensionality (up to logarithms). As in the full grid case, let $T = (0,1)^m$ and let $\rho_T = \lambda_T$ be the Lebesgue measure. However, we now assume that the true solution comes from a mixed Sobolev space, i.e. let $f_\rho \in H^s_{\text{mix}}(T;\mathbb{R}^d)$ for a $0 < s \leq 2$. We choose $V_k = \mathcal{V}_k^{\text{sparse},d}$ from (3.34). Note that - as in the full grid case - a properly adjusted sparse grid space might

be more appropriate if the regression function is known to belong to $H^{\hat{s}}_{\text{mix}}(T;\mathbb{R}^d)$ with $\hat{s} > 2$ or to a mixed smoothness class of varying degrees for different directions, see [39]. Furthermore, adaptive sparse grids can be employed to deal with non-smooth solutions, see e.g. [15]. An exhaustive analysis of appropriate sparse grid spaces V_k is beyond the scope of this thesis, but we refer the reader to [14] and [52] for details in this direction.

We choose $\|\cdot\|_{V_k} = \|\cdot\|_{H^1_{\text{mix}}(T;\mathbb{R}^d)}$. As we already mentioned in subsection 3.3.2, the mixed space $H^1_{\text{mix}}(T;\mathbb{R}^d)$ is a reproducing kernel Hilbert space for which proposition 3.18 holds for every dimension m. Therefore, $c_{V_k} \simeq 1$ can be chosen independently of $k \in \mathbb{N}$.

Note that, in the case $m = 1$, a sparse grid coincides with a full grid and $H^s_{\text{mix}}(T;\mathbb{R}^d) = H^s(T;\mathbb{R}^d)$. Hence, our following analysis complements the full grid study from the last subsection, where we omitted the case $m = 1$.

The overall error

In (3.40), we provided the inverse inequality

$$\|f\|_{H^1_{\text{mix}}(T;\mathbb{R}^d)} \leq \tilde{c}2^k \|f\|_{L_2(T;\mathbb{R}^d)} \quad \forall\, f \in V_k$$

with m- and d-dependent constant $\tilde{c} > 0$. Therefore, we can apply theorem 4.18 with $b := \tilde{c}2^k r \geq \tilde{c}2^k \|f_\rho\|_{L_2(T;\mathbb{R}^d)}$ to obtain

$$\mathcal{E}(f_{V_{k,b}}) - \mathcal{E}(f_\rho) = \inf_{f \in V_k} \|f - f_\rho\|^2_{L_2(T;\mathbb{R}^d)} \overset{(3.42)}{=} \mathcal{O}\left(2^{-2sk} k^{m-1}\right)$$

for sparse grids. Since $c_{V_k} \simeq 1$, $b \simeq 2^k$ and $N_k \simeq 2^k k^{m-1}$, we derive

$$\begin{aligned}\mathcal{E}(f_{\mathbf{z}_n,V_{k,b}}) - \mathcal{E}(f_\rho) &\lesssim 2^{-2sk} k^{m-1} + \frac{M_\psi N_k}{n} \max\left(1, \log\left(\frac{c_{V_k} bn}{\delta\sqrt{M_\psi N_k}}\right)\right) \\ &\lesssim 2^{-2sk} k^{m-1} + \frac{M_\psi 2^k k^{m-1}}{n} \max\left(1, \log\left(\frac{n}{\delta\sqrt{M_\psi} k^{m-1}}\right)\right), \qquad (4.51)\end{aligned}$$

with confidence $1 - \delta$ for the overall error by combining the result on the discretization error with the sampling error bound from theorem 4.24. In the following, we will discern the case $M_\psi \simeq 1$ and the case $M_\psi \simeq (c_{V_k} b + r)^2 \simeq b^2 \simeq 2^{2k}$, where we used $c_{V_k} \simeq 1$.

The convergent case

Now, let us consider (4.51) in more detail. Similar to the full grid case, we obtain that $n > 2^k k^{m-1} \simeq N_k$ is a necessary condition for convergence and that the maximum function evaluates to its second argument regardless of the choice of M_ψ. Thus, we derive the rate

$$\mathcal{E}(f_{\mathbf{z}_n,V_{k,b}}) - \mathcal{E}(f_\rho) \lesssim 2^{-2sk} k^{m-1} + \frac{2^k k^{m-1}}{n} \log\left(\frac{n}{\delta k^{m-1}}\right) \qquad (4.52)$$

for constant M_ψ and

$$\mathcal{E}(f_{\mathcal{Z}_n,V_{k,b}}) - \mathcal{E}(f_\rho) \lesssim 2^{-2sk}k^{m-1} + \frac{2^{3k}k^{m-1}}{n}\log\left(\frac{n}{\delta 2^k k^{m-1}}\right) \tag{4.53}$$

for $M_\psi \simeq 2^{2k}$. Furthermore, we observe that

$$N_k \ll \frac{n}{\log(n)}$$

is a sufficient condition for convergence in the first case $M_\psi \simeq 1$ and

$$N_k^3 \ll \frac{n}{\log(n)}$$

is a sufficient condition for convergence in the second case $M_\psi \simeq 2^{2k}$.

Balancing the error terms

We first examine the case where M_ψ is constant and equate

$$2^{-2sk}k^{m-1} \simeq \frac{2^k k^{m-1}}{n}\log\left(\frac{n}{\delta k^{m-1}}\right),$$

which leads to

$$n \simeq 2^{(2s+1)k}(\log(n) - (m-1)\log(k)).$$

Thus, up to logarithmic factors in N_k and n, the optimal scaling is

$$n \simeq 2^{(2s+1)k}.$$

Note that this also implies $\log(n) \simeq k$ for $n > 1$. Substituting this relation into (4.52), the overall rate can be bounded by

$$\begin{aligned}\mathcal{E}(f_{\mathcal{Z}_n,V_{k,b}}) - \mathcal{E}(f_\rho) &\lesssim n^{-\frac{2s}{2s+1}}\left(\frac{1}{2s+1}\log(n)\right)^{m-1} \\ &\quad\cdot\left(1 + \log(n) - \log(\delta) - (m-1)\log\left(\frac{1}{2s+1}\log(n)\right)\right) \\ &= \mathcal{O}\left(n^{-\frac{2s}{2s+1}}\log(n)^m\right).\end{aligned}$$

In the case $M_\psi \simeq (c_{V_k}b + r)^2 \simeq 2^{2k}$, we have to balance

$$2^{-2sk}k^{m-1} \simeq \frac{2^{3k}k^{m-1}}{n}\log\left(\frac{n}{\delta 2^k k^{m-1}}\right)$$

and get

$$n \simeq 2^{(3+2s)k} \left(\log(n) - k\log(2) - (m-1)\log(k)\right),$$

which essentially is

$$n \simeq 2^{(3+2s)k}$$

up to logarithms. Employing this in (4.53), we obtain

$$\begin{aligned}\mathcal{E}(f_{\mathcal{Z}_n,V_{k,b}}) - \mathcal{E}(f_\rho) &\lesssim n^{-\frac{2s}{2s+3}} \left(\frac{1}{2s+3}\log(n)\right)^{m-1} \\ &\quad \cdot \left(1 + \log(n) - \log(\delta) - \frac{1}{2s+3}\log(n) - (m-1)\log\left(\frac{1}{2s+3}\log(n)\right)\right) \\ &= \mathcal{O}\left(n^{-\frac{2s}{2s+3}}\log(n)^m\right).\end{aligned}$$

4.5.3 Periodic regression with Fourier polynomials on full grids

In the following, we consider regression of multivariate periodic functions. In this subsection we focus on Fourier polynomials on full grids, see also subsection 3.5.2. To this end, let $T = (-\pi,\pi)^m$ and let $\rho_T = \frac{1}{(2\pi)^m}\lambda_T$ be the rescaled Lebesgue measure. Let, furthermore, $f_\rho \in \bar{H}^s(T;\mathbb{C}^d)$ stem from a periodic Sobolev space[6] of degree $s > 0$. In contrast to the previous examples, we can now also exploit the case of higher smoothness $s > 2$. We choose $V_k = \mathcal{T}_k^{\text{full},d}$ as the space of Fourier polynomials on a full frequency grid of level $k > 0$, see (3.43). Therefore, $N_k = \dim(V_k) \simeq 2^{km}$.

As we mentioned above, we deliberately choose $\|\cdot\|_{V_k} = \|\cdot\|_{\bar{H}^1(T;\mathbb{C}^d)}$ as regularization norm. However, due to the smoothness of the Fourier polynomials, we could also consider higher-order Sobolev norms. Similarly to the non-periodic case, the space $\bar{H}^1(T;\mathbb{C}^d)$ is not a reproducing kernel Hilbert space for $m \geq 2$. Therefore, we cannot assume that the embedding constant c_{V_k} is bounded independently of $k \in \mathbb{N}$ for $m \geq 2$. Again, we lean on a Nikolskii-type inequality to obtain

$$\|f\|_\infty \lesssim 2^{\frac{km}{2}}\|f\|_{L_2(T;\mathbb{C}^d)} \lesssim 2^{\frac{km}{2}}\|f\|_{\bar{H}^1(T;\mathbb{C}^d)} \quad \forall\, f \in V_k$$

in this case, see theorem II.2.2 of [78]. Thus, we have $c_{V_k} \simeq 1$ for $m = 1$ and we assume $c_{V_k} \simeq 2^{\frac{km}{2}}$ for $m \geq 2$. In the following, we will omit the analysis of the special case $m = 1$ because it will coincide with the case $m = 1$ for the hyperbolic cross example in the next section.

[6] Note that we use $\bar{H}^s(T;\mathbb{C}^d)$ instead of the probability space $\bar{H}^s_{2,\rho_T}(T;\mathbb{C}^d)$. However, both spaces contain the same elements and their norms only differ by the constant factor $(2\pi)^m$. Since we are only interested in convergence rates with respect to the grid level k and the number of samples n, we can neglect this prefactor.

The overall error

Let $m \geq 2$. Because of the inverse inequality

$$\|f\|_{\bar{H}^1(T;\mathbb{C}^d)} \leq \tilde{c}2^k \|f\|_{L_2(T;\mathbb{C}^d)} \quad \forall\, f \in V_k$$

with constant $\tilde{c} > 0$, see (3.48), the application of theorem 4.18 with $b := \tilde{c}2^k r \geq \tilde{c}2^k \|f_\rho\|_{L_2(T;\mathbb{C}^d)}$ leads to

$$\mathcal{E}(f_{V_{k,b}}) - \mathcal{E}(f_\rho) = \inf_{f \in V_k} \|f - f_\rho\|^2_{L_2(T;\mathbb{C}^d)} \overset{(3.50)}{=} \mathcal{O}\left(2^{-2sk}\right).$$

Combining this result with the sampling error rate from theorem 4.24, we obtain the bound

$$\begin{aligned}\mathcal{E}(f_{\mathcal{Z}_n,V_{k,b}}) - \mathcal{E}(f_\rho) &\lesssim 2^{-2sk} + \frac{M_\psi N_k}{n} \max\left(1, \log\left(\frac{c_{V_k} bn}{\delta\sqrt{M_\psi} N_k}\right)\right) \\ &\lesssim 2^{-2sk} + \frac{M_\psi 2^{km}}{n} \max\left(1, \log\left(\frac{n}{\delta\sqrt{M_\psi}\sqrt{2^{k(m-2)}}}\right)\right),\end{aligned}$$

with confidence $1 - \delta$ for $m \geq 2$, $c_{V_k} \simeq 2^{\frac{km}{2}}$, $b \simeq 2^k$ and $N_k \simeq 2^{km}$.

The convergent case

Since $n > 2^{km} \simeq N_k$ is again a necessary condition for convergence of the error to 0, the rate becomes

$$\mathcal{E}(f_{\mathcal{Z}_n,V_{k,b}}) - \mathcal{E}(f_\rho) \lesssim 2^{-2sk} + \frac{2^{km}}{n} \log\left(\frac{n}{\delta\sqrt{2^{k(m-2)}}}\right) \tag{4.54}$$

for $M_\psi = (M + r)^2 \simeq 1$ and

$$\mathcal{E}(f_{\mathcal{Z}_n,V_{k,b}}) - \mathcal{E}(f_\rho) \lesssim 2^{-2sk} + \frac{2^{2k(m+1)}}{n} \log\left(\frac{n}{\delta 2^{km}}\right) \tag{4.55}$$

for $M_\psi = (c_{V_k} b + r)^2 \simeq 2^{k(m+2)}$. Since $m \geq 2$, a sufficient condition for convergence of the error to 0 is

$$N_k \ll \frac{n}{\log(n)}$$

if $M_\psi \simeq 1$. In the case $M_\psi \simeq 2^{k(m+2)}$, a sufficient condition is given by

$$\left(N_k 2^k\right)^2 \ll \frac{n}{\log(n)}.$$

Note that the number of samples n suffers from the curse of dimensionality since - regardless of the choice of M_ψ - it has to grow faster than $N_k \simeq 2^{km}$.

Balancing the error terms

Equating the discretization error and the sampling error in (4.54), we obtain

$$2^{-2sk} \simeq \frac{2^{km}}{n} \log\left(\frac{n}{\delta\sqrt{2^{k(m-2)}}}\right) \;\Leftrightarrow\; n \simeq 2^{(2s+m)k} \log\left(\frac{n}{\delta\sqrt{2^{k(m-2)}}}\right)$$

and, thus, $n \simeq 2^{(2s+m)k}$ as the optimal coupling up to logarithms in n and N_k. Therefore, for a fixed confidence $1-\delta$, we derive

$$\begin{aligned}\mathcal{E}(f_{\mathcal{Z}_n,V_{k,b}}) - \mathcal{E}(f_\rho) &\lesssim n^{-\frac{2s}{2s+m}} \left(1 + \frac{2s+\frac{m}{2}+1}{2s+m}\log(n) - \log(\delta)\right) \\ &= \mathcal{O}\left(n^{-\frac{2s}{2s+m}}\log(n)\right)\end{aligned}$$

as the error rate for the optimal coupling if $M_\psi \simeq 1$. If $M_\psi \simeq (c_{V_k} b + r)^2 \simeq 2^{k(m+2)}$, we get

$$2^{-2sk} \simeq \frac{2^{2k(m+1)}}{n} \log\left(\frac{n}{\delta 2^{km}}\right) \;\Leftrightarrow\; n \simeq 2^{2k(m+s+1)} \log\left(\frac{n}{\delta 2^{km}}\right)$$

by balancing the errors in (4.55) and, therefore, $n \simeq 2^{2k(m+s+1)}$ up to logarithms. This scaling leads to the rate

$$\begin{aligned}\mathcal{E}(f_{\mathcal{Z}_n,V_{k,b}}) - \mathcal{E}(f_\rho) &\lesssim n^{-\frac{s}{s+m+1}} \left(1 + \frac{2s+m+2}{2(s+m+1)}\log(n) - \log(\delta)\right) \\ &= \mathcal{O}\left(n^{-\frac{s}{s+m+1}}\log(n)\right).\end{aligned}$$

4.5.4 Periodic regression with Fourier polynomials on hyperbolic crosses

Finally, we examine regression of multivariate periodic functions by Fourier polynomials on hyperbolic crosses, see subsection 3.5.2. As in the last subsection, let $T = (-\pi,\pi)^m$ and let $\rho_T = \frac{1}{(2\pi)^m}\lambda_T$. Let $f_\rho \in \bar{H}^s_{\text{mix}}(T;\mathbb{C}^d)$ reside in the mixed Sobolev space of smoothness $s > 0$. Our search sets are $V_k = \mathcal{T}_k^{\text{hyp},d}$ for $k \in \mathbb{N}$ with $N_k \simeq 2^k k^{m-1}$, cf. (3.44) and (3.45).

We take $\|\cdot\|_{V_k} = \|\cdot\|_{\bar{H}^1_{\text{mix}}(T;\mathbb{C}^d)}$ as regularization norm, which allows us to choose $c_{V_k} \simeq 1$ independently of k since $\bar{H}^1_{\text{mix}}(T;\mathbb{C}^d)$ is a reproducing kernel Hilbert space for which proposition 3.18 applies. In the case $m = 1$, our setting is exactly the same as for full grids. Therefore, our results here also complement our earlier full grid analysis, where we omitted the case $m = 1$.

The overall error

Taking the inverse inequality

$$\|f\|_{\bar{H}^1_{\mathrm{mix}}(T;\mathbb{C}^d)} \leq \tilde{c}2^k \|f\|_{L_2(T;\mathbb{C}^d)} \quad \forall\, f \in V_k,$$

with a constant $\tilde{c} > 0$, into account, see (3.49), we obtain that $b := \tilde{c}2^k r \geq \tilde{c}2^k \|f_\rho\|_{L_2(T;\mathbb{C}^d)}$ is a valid choice in theorem 4.18. Therefore, we get

$$\mathcal{E}(f_{V_{k,b}}) - \mathcal{E}(f_\rho) = \inf_{f \in V_k} \|f - f_\rho\|^2_{L_2(T;\mathbb{C}^d)} \overset{(3.51)}{=} \mathcal{O}\left(2^{-2sk}\right).$$

Using $c_{V_k} \simeq 1$, $b \simeq 2^k$ and $N_k \simeq 2^k k^{m-1}$, we derive

$$\begin{aligned}\mathcal{E}(f_{\mathcal{Z}_n,V_{k,b}}) - \mathcal{E}(f_\rho) &\lesssim 2^{-2sk} + \frac{M_\psi N_k}{n} \max\left(1, \log\left(\frac{c_{V_k} bn}{\delta\sqrt{M_\psi N_k}}\right)\right) \\ &\lesssim 2^{-2sk} + \frac{M_\psi 2^k k^{m-1}}{n} \max\left(1, \log\left(\frac{n}{\delta\sqrt{M_\psi} k^{m-1}}\right)\right) \end{aligned} \tag{4.56}$$

with confidence $1 - \delta$ by summing up the discretization error and the sampling error from theorem 4.24.

The convergent case

Since $n > 2^k k^{m-1} \simeq N_k$ is a necessary condition for convergence of (4.56) to 0 for $k, n \to \infty$, we observe that

$$\mathcal{E}(f_{\mathcal{Z}_n,V_{k,b}}) - \mathcal{E}(f_\rho) \lesssim 2^{-2sk} + \frac{2^k k^{m-1}}{n} \log\left(\frac{n}{\delta k^{m-1}}\right) \tag{4.57}$$

for $M_\psi \simeq 1$ and

$$\mathcal{E}(f_{\mathcal{Z}_n,V_{k,b}}) - \mathcal{E}(f_\rho) \lesssim 2^{-2sk} + \frac{2^{3k} k^{m-1}}{n} \log\left(\frac{n}{\delta 2^k k^{m-1}}\right) \tag{4.58}$$

for $M_\psi \simeq (c_{V_k} b + r)^2 \simeq b^2 \simeq 2^{2k}$ are the error bounds in the convergent case. Since the sampling error term reads the same as for sparse grids in subsection 4.5.2, the sufficient conditions for the error convergence can be inherited from there. We deduce

$$N_k \ll \frac{n}{\log(n)}$$

if $M_\psi \simeq 1$. For the case $M_\psi \simeq 2^{2k}$, the condition becomes

$$N_k^3 \ll \frac{n}{\log(n)}.$$

Balancing the error terms

First, let us consider the case $M_\psi \simeq 1$. To this end, we equate

$$2^{-2sk} \simeq \frac{2^k k^{m-1}}{n} \log\left(\frac{n}{\delta k^{m-1}}\right) \;\Leftrightarrow\; n \simeq 2^{(2s+1)k} k^{m-1} \log\left(\frac{n}{\delta k^{m-1}}\right).$$

Thus, the optimal scaling up to logarithmic factors in the basis size N_k and the sample size n is $n \simeq 2^{(2s+1)k}$. Rewriting (4.57) with the help of this relation gives

$$\begin{aligned}\mathcal{E}(f_{\mathcal{Z}_n,V_{k,b}}) - \mathcal{E}(f_\rho) &\lesssim n^{-\frac{2s}{2s+1}} + n^{-\frac{2s}{2s+1}} \left(\frac{1}{2s+1}\log(n)\right)^{m-1} \\ &\quad \cdot \left(\log(n) - \log(\delta) - (m-1)\log\left(\frac{1}{2s+1}\log(n)\right)\right) \\ &= \mathcal{O}\left(n^{-\frac{2s}{2s+1}} \log(n)^m\right)\end{aligned}$$

for $n \to \infty$. Second, we have a look at the case $M_\psi \simeq 2^{2k}$ and consider the coupling

$$2^{-2sk} \simeq \frac{2^{3k} k^{m-1}}{n} \log\left(\frac{n}{\delta 2^k k^{m-1}}\right) \;\Leftrightarrow\; n \simeq 2^{(2s+3)k} k^{m-1} \log\left(\frac{n}{\delta 2^k k^{m-1}}\right),$$

which results in $n \simeq 2^{(2s+3)k}$ up to logarithms. Substituting this into (4.58), we obtain

$$\begin{aligned}\mathcal{E}(f_{\mathcal{Z}_n,V_{k,b}}) - \mathcal{E}(f_\rho) &\lesssim n^{-\frac{2s}{2s+3}} + n^{-\frac{2s}{2s+3}} \left(\frac{1}{2s+3}\log(n)\right)^{m-1} \\ &\quad \cdot \left(\log(n) - \log(\delta) - \frac{1}{2s+3}\log(n) - (m-1)\log\left(\frac{1}{2s+3}\log(n)\right)\right) \\ &= \mathcal{O}\left(n^{-\frac{2s}{2s+3}} \log(n)^m\right).\end{aligned}$$

4.5.5 Overview

An overview on the results for our examples can be found in table 4.2. Note that the full grid results are valid only if $m \geq 2$. Note, furthermore, that the results for Fourier polynomials on hyperbolic crosses read exactly the same as the results for piecewise linear prewavelets on sparse grids. The only difference is that the smoothness s is constrained to $0 < s \leq 2$ in the latter case because the piecewise linear prewavelets cannot exploit higher orders of Sobolev smoothness of f_ρ.

(a) Piecewise linear prewavelets on full grids and and sparse grids, $0 < s \leq 2$

	V_k	suff. cond.	balanced n	balanced rate
$M_\psi \simeq 1$	$\mathcal{V}_k^{\text{full},d}$	$N_k \ll \frac{n}{\log(n)}$	$n \simeq 2^{(2s+m)k}$	$n^{-\frac{2s}{2s+m}} \log(n)$
	$\mathcal{V}_k^{\text{sparse},d}$	$N_k \ll \frac{n}{\log(n)}$	$n \simeq 2^{(2s+1)k}$	$n^{-\frac{2s}{2s+1}} \log(n)^m$
$M_\psi \simeq (c_{V_k} b + r)^2$	$\mathcal{V}_k^{\text{full},d}$	$2^{2k} N_k^3 \ll \frac{n}{\log(n)}$	$n \simeq 2^{(2s+3m+2)k}$	$n^{-\frac{2s}{2s+3m+2}} \log(n)$
	$\mathcal{V}_k^{\text{sparse},d}$	$N_k^3 \ll \frac{n}{\log(n)}$	$n \simeq 2^{(2s+3)k}$	$n^{-\frac{2s}{2s+3}} \log(n)^m$

(b) Fourier polynomials on full grids and hyperbolic crosses, $s > 0$

	V_k	suff. cond.	balanced n	balanced rate
$M_\psi \simeq 1$	$\mathcal{T}_k^{\text{full},d}$	$N_k \ll \frac{n}{\log(n)}$	$n \simeq 2^{(2s+m)k}$	$n^{-\frac{2s}{2s+m}} \log(n)$
	$\mathcal{T}_k^{\text{hyp},d}$	$N_k \ll \frac{n}{\log(n)}$	$n \simeq 2^{(2s+1)k}$	$n^{-\frac{2s}{2s+1}} \log(n)^m$
$M_\psi \simeq (c_{V_k} b + r)^2$	$\mathcal{T}_k^{\text{full},d}$	$2^{2k} N_k^2 \ll \frac{n}{\log(n)}$	$n \simeq 2^{(2s+2m+2)k}$	$n^{-\frac{s}{s+m+1}} \log(n)$
	$\mathcal{T}_k^{\text{hyp},d}$	$N_k^3 \ll \frac{n}{\log(n)}$	$n \simeq 2^{(2s+3)k}$	$n^{-\frac{2s}{2s+3}} \log(n)^m$

Table 4.2: Results for constrained regression (B) over $V_{k,b}$ for full grids ($m \geq 2$, H^1 regularization, $f_\rho \in H^s$, $N_k \simeq 2^{km}$) and for sparse grids/hyperbolic crosses ($m \geq 1$, H^1_{mix} regularization, $f_\rho \in H^s_{\text{mix}}$, $N_k \simeq 2^k k^{m-1}$). The scaling of the regularization parameter is $b \simeq 2^k$ in all cases. The table contains a sufficient condition for convergence of the overall error to 0, the behavior of n when the discretization error and the sampling error are balanced (up to logarithms in n and N_k) and the corresponding convergence rate in the balanced case.

The curse of dimensionality

For full grids, we directly observe how the curse of dimensionality affects both the number of sample points n and the convergence rate of the regression error: Our conditions for convergence imply that the sample size n has to grow faster than the size N_k of the finite-dimensional search space V_k in all analyzed cases. In the full grid case, this essentially means that we need at least $N_k \simeq 2^{mk}$ sample points. Here, the exponential dependence with respect to m shows up. Another way to see the presence of the curse of dimensionality is to investigate the convergence rate in the balanced case. Take for instance $M_\psi \simeq 1$. There, the rate of convergence in the balanced case reads $n^{-\frac{2s}{2s+m}} \log(n)$ for full grids. Therefore, the sample size has to grow exponentially with respect to m in order to achieve the same convergence rate as in the univariate case.

In the case of a sparse grid or hyperbolic cross discretization, the curse of dimensionality is present only in a weak form since it affects the logarithms in n and N_k. To this end, note that the condition on the number of sampling points n is again stated with respect to the basis size N_k, which scales like $2^k k^{m-1}$ for sparse grids and hyperbolic crosses. Therefore, even in the most restrictive case $N_k^3 \ll \frac{n}{\log(n)}$ for $M_\psi = (c_{V_k} b + r)^2$,

the condition for convergence essentially reads $2^{3k}k^{3(m-1)} \ll \frac{n}{\log(n)}$ and the curse of dimensionality only affects the level k. Similarly, for the convergence rate in the balanced case, the dimension m is present only in the factor $\log(n)^m$ but not in the main term, as it is the case for full grids.

Oversampling in the optimal/balanced case in dependence on M_ψ

To achieve the optimal rate of convergence, we calculated the amount n of samples (up to constants and logarithms) such that the discretization error and the sampling error are approximately equal. We now want to illustrate how M_ψ influences the amount of samples in the balanced case. To this end, we consider the linear prewavelet basis. Note, however, that an analogous analysis can also be done for Fourier polynomials.

Since we are interested in a result which relates the basis size N_k to the sample size n, we express the number of samples in the balanced case (up to logarithms) in terms of N_k. For $M_\psi \simeq 1$, we obtain

$$n \simeq 2^{(2s+m)k} = 2^{(2\frac{s}{m}+1)mk} \simeq N_k^{\frac{2s}{m}+1}$$

for the full grid space $\mathcal{V}_k^{\text{full},d}$ with $N_k \simeq 2^{mk}$ and

$$n \simeq 2^{(2s+1)k} \lesssim N_k^{2s+1}$$

for the sparse grid space $\mathcal{V}_k^{\text{sparse},d}$ with $N_k \simeq 2^k k^{m-1}$. Therefore, the larger the smoothness index s is, the higher the oversampling has to be in terms of a power of the basis size N_k. Note that for growing s, the oversampling power $2s+1$ for sparse grids grows faster than the oversampling power $\frac{2s}{m}+1$ for full grids. This, however, is of course only with respect to the particular basis size. For s close to 0, the sampling relation approaches $n \simeq N_k$ for full grids and sparse grids, which essentially reflects the convergence condition $N_k \ll \frac{n}{\log(n)}$ from table 4.2 (up to logarithms).

Now, we investigate the case $M_\psi = (c_{V_k} b + r)^2$, for which we have

$$n \simeq 2^{(2s+3m+2)k} \simeq 2^{(\frac{2s}{m}+3+\frac{2}{m})mk} \simeq N_k^{\frac{2s+2}{m}+3} \simeq N_k^{\frac{2s}{m}+1} N_k^{\frac{2}{m}+2}$$

in the balanced case for the full grid space $\mathcal{V}_k^{\text{full},d}$. Analogously, we have

$$n \simeq 2^{(2s+3)k} \lesssim N_k^{2s+3} \simeq N_k^{2s+1} N_k^2$$

for the sparse grid space $\mathcal{V}_k^{\text{sparse},d}$. The qualitative behavior is the same as in the case $M_\psi \simeq 1$. However, regardless of $s > 0$, the optimal number of sample points n is always larger than $N_k^{\frac{2}{m}+3} \simeq 2^{2k} N_k^3$ for full grids and larger than N_k^3 for sparse grids. This again represents the sufficient conditions for convergence of the error to 0, which we provided in table 4.2. Generally, we see that the main difference in the oversampling for

$M_\psi = (c_{V_k} b + r)^2$ in contrast to $M_\psi \simeq 1$ is that we need an additional factor of at least N_k^2 here. Depending on the level k and the dimension m, this can be a huge number.

As we mentioned in subsection 4.4.1, it is usually not easy to prove that one can work with $M_\psi = (M+r)^2 \simeq 1$ instead of $M_\psi = (c_{V_k} b + r)^2$. However, in many applications the assumption $M_\psi \simeq 1$ is reasonable since the L_∞ norm of all functions under consideration is bounded by an absolute constant. Our results in table 4.2 and the reasoning above shows that the actual behavior of M_ψ has a serious influence not only on the convergence rate but also on the required oversampling.

Convergence with respect to n

The fastest convergence can be observed in the case of hyperbolic cross regression with $M_\psi \simeq 1$, where we have the rate

$$n^{-\frac{2s}{2s+1}} \log(n)^m$$

in the balanced case. Assuming that f_ρ is smooth, i.e. $f_\rho \in H^s_{\mathrm{mix}}(T; \mathbb{C}^d)$ for all $s > 0$, we obtain that the convergence rate is essentially n^{-1} up to logarithms since $\frac{2s}{2s+1} \to 1$ for $s \to \infty$. Note that this is also the best we could expect for any example. This is due to the fact that our results rely on theorem 4.24, where the decay of the sampling error cannot exceed the rate n^{-1}.

The main reason why we cannot derive improved results on the convergence of the sampling error is the fact that we stated the regression problem (A) in a very general manner. We did not pose restrictions on the measure ρ and we allowed for noisy samples for instance. In the special case of noiseless function regression, however, we can achieve better convergence rates than n^{-1}. We discuss this in detail in section 5.4.

4.5.6 Relation to other results

In this subsection, we relate our results to the work of other researchers.

Convergence conditions and stability of unconstrained regression

First, we observe that our sufficient conditions for convergence directly correspond to certain stability conditions for unregularized regression. To this end, recall that we enforced stability and well-posedness of the regression problem by introducing a constraint on $\|\cdot\|_{V_k}$. When omitting this constraint, one has to identify a coupling between n and N_k for which stability and well-posedness of the regression problem is guaranteed with high probability before convergence of the error can be obtained. This is essentially the difference between well-posedness for a Tikhonov-regularized problem and for a discretized problem without penalty term, see also [35] for a more detailed explanation.

Using a truncation operator, which enforces a similar condition as our assumption $M_\psi \simeq 1$, such couplings between n and N_k are determined e.g. in [19, 21, 61] for search

spaces of global polynomials. Furthermore, in [21], we also find a result which states that

$$N_k \leq \frac{n}{\log(n)}$$

is sufficient to get a well-posed and stable regression problem for orthonormal piecewise constant basis functions, which can easily be extended to higher order bases. This stability condition matches our sufficient conditions for convergence of prewavelet and Fourier polynomial regression for $M_\psi \simeq 1$, see table 4.2. Note, however, that the results in [21] are valid only if the considered basis is orthonormal. We will discuss this issue in more detail in the next chapter.

Regression on full grids with linear splines

Next, we have a look at a standard result from non-parametric regression on full grids: Omitting the H^1 regularization and considering a truncated variant of the scalar-valued regression problem, the convergence rate for multivariate regression with splines on a full grid has been provided in e.g. theorem 15.4 of [46]. There, the bound

$$2^{-2sk} + \frac{2^{km}\log(n)}{n}$$

for $0 < s \leq 2$ is obtained in the piecewise linear case, which reflects the rate[7] we have shown in (4.46). For a specific setting where the degrees of freedom N_k are coupled to the smoothness s, they also obtained $n^{-\frac{2s}{2s+m}}$ as the rate in the balanced case up to logarithms. This coincides with the rate which we observed in (4.49).

Regression on sparse grids with linear splines

In [36], the limit behavior of the dual regression problem, which we consider in the next chapter, is studied for the application of the so-called combination technique for linear splines on sparse grids. There, the solution to the regression problem is computed on smaller full grids and then combined linearly to obtain an approximate solution on the sparse grid, see also [35]. Note that this method does not necessarily lead to the true sparse grid solution of the regression problem. Note, furthermore, that [36] considers H^1 regularization instead of H^1_{mix}. Besides a thorough analysis in the case of fixed n, the authors also give a conjecture on the overall error bound in the H^1 norm, which can be reformulated as

$$\|f_\rho - f^{\text{sol}}\|_{H^1(T)} \lesssim \inf_{f \in V_k} \|f_\rho - f\|_{H^1(T)} + \frac{\sqrt{\text{dof}(f^{\text{sol}})}}{\sqrt{n}},$$

[7] Note, that we have proven convergence in probability, whereas the statement in [46] is with respect to $L_2(T)$ convergence. However, since $\mathcal{E}(f_{\mathcal{Z}_n, V_{k,b}})$ is bounded from above for all $n \in \mathbb{N}$, these notions of convergence are equivalent.

where f^{sol} is the combination technique solution, which approximates $f_{\mathcal{Z}_n,V_{k,b}}$, and $\text{dof}(f^{\text{sol}})$ is the sum of the number of degrees of freedom which are employed on the full grids to obtain the combined sparse grid solution. Apart from the fact that the H^1 norm is estimated instead of the regression error (4.24), which corresponds to the L_2 norm, (4.52) can be seen as the analogous bound when directly working on the sparse grid instead of considering the combination technique. Note that our estimate is for the squared L_2 norm, see lemma 4.4, and we have to take the square root of the rate in (4.24). Then our rate reflects the conjecture from [36] (up to logarithms) when substituting the H^1 norm by the L_2 norm. However, there might still be room for improvement since the conjecture from [36] is essentially based on a central limit theorem estimate, which could be improved for convex search sets by the techniques we have provided in section 4.4.

A different approach with Jackson and Bernstein inequalities

The authors of [75] present a result which is also based on Jackson and Bernstein inequalities to obtain an upper bound on the discretization error for regularized regression. Therefore, their work comes close to the situation we have been dealing with in this chapter. However, there are slight differences in their prerequisites and their results compared to what we presented. We now state a version of their result which is adapted to our notation.

Theorem 4.25 [Upper bound for the discretization error from [75]]
Let $C > 0$, let E be a separable Hilbert space and let $(\tau_k)_{k\in\mathbb{N}}$ be a decreasing sequence of positive numbers which converges to 0. Let $(\mathcal{H}, \|\cdot\|_{\mathcal{H}})$ be an infinite-dimensional Banach space with continuous embedding $\mathcal{H} \hookrightarrow L_2(T;E)$ and let the finite-dimensional search spaces $V_k \subset \mathcal{H}$ be normed by $\|\cdot\|_{V_k} = \|\cdot\|_{\mathcal{H}}$ for $k \in \mathbb{N}$. Furthermore, let $\mathcal{P}_k : L_2(T;E) \to V_k$ be linear operators with $\|\mathcal{P}_k\|_{\mathcal{L}(L_2(T;E),L_2(T;E))} \leq C$ for all $k \in \mathbb{N}$. We assume that the Jackson inequalities

$$\|\mathcal{P}_k(f) - f\|_{L_2(T;E)} \leq C\tau_{k+1}\|f\|_{\mathcal{H}} \;\forall\, f \in \mathcal{H}$$

and the Bernstein inequalities

$$\|\mathcal{P}_k(f)\|_{\mathcal{H}} \leq C\tau_k^{-1}\|f\|_{L_2(T;E)} \;\forall\, f \in L_2(T;E) \quad \text{and} \quad \|\mathcal{P}_k(f)\|_{\mathcal{H}} \leq C\|f\|_{\mathcal{H}} \;\forall\, f \in \mathcal{H}$$

hold for all $k \in \mathbb{N}$. If $f_\rho \in (L_2(T;E), \mathcal{H})_\theta$ for some $0 < \theta < 1$ and $b = C\tau_k^{\theta-1}\|f_\rho\|_\theta$, the discretization error is bounded by

$$\mathcal{E}(f_{V_{k,b}}) - \mathcal{E}(f_\rho) \lesssim b^{-\frac{2\theta}{1-\theta}} \simeq \tau_k^{2\theta}.$$

Proof. See theorem 4.1 of [75]. □

First of all, we observe that the growth of b in theorem 4.25 is now coupled also to the smoothness, which is implicitly given by the interpolation parameter θ. For θ close

to 1, the growth of b - and therefore also the regularization - is very mild. This is a clear advantage in contrast to the choice from lemma 4.17, where we had to rely on the Bernstein inequality with respect to the V_k norm to obtain b. However, the main drawback in theorem 4.25 is that the V_k norm has to be the norm of the space for which the Jackson inequalities are valid.

To illustrate the application of the above theorem, we consider regression with Fourier polynomials on hyperbolic crosses as in subsection 4.5.4. Since the $\mathcal{H}$ norm in theorem 4.25 represents the regularization norm, we choose $\mathcal{H} = \bar{H}^1_{\text{mix}}(T;\mathbb{C}^d)$ to be consistent with our example. If $\mathcal{P}_k$ is chosen to be the $L_2(T;\mathbb{C}^d)$-orthogonal projection P_{V_k}, see definition 4.16, the prerequisites of theorem 4.25 are fulfilled for $\tau_k = 2^{-k}$, cf. (3.49), (3.51) and [78] for details. Therefore, we obtain

$$\mathcal{E}(f_{V_{k,b}}) - \mathcal{E}(f_\rho) \lesssim \tau_k^{2\theta} = 2^{-2\theta k}$$

if we choose $b = C\tau_k^{\theta-1}\|f_\rho\|_\theta \simeq 2^{k(1-\theta)}$ and if $f_\rho \in \left(L_2(T;\mathbb{C}^d), \bar{H}^1_{\text{mix}}(T;\mathbb{C}^d)\right)_\theta$ for some $\theta \in (0,1)$. Clearly, the coupling $b \simeq 2^{k(1-\theta)}$ is beneficial in contrast to $b \simeq 2^k$, which we had in subsection 4.5.4. However, because of the choice $\mathcal{H} = \bar{H}^1_{\text{mix}}(T;\mathbb{C}^d)$, the decay rate of the discretization error cannot be better than 2^{-2k} in contrast to subsection 4.5.4, where we were able to exploit additional smoothness of f_ρ and obtain a rate of 2^{-2sk} for $f_\rho \in \bar{H}^s_{\text{mix}}(T;\mathbb{C}^d)$ with $s > 0$. To exploit this additional smoothness, we would have to change the regularization norm from the mixed Sobolev norm for smoothness index 1 to the mixed Sobolev norm for smoothness index s, which does not correspond to the setting from subsection 4.5.4 anymore.

When considering the application of theorem 4.25 to the sparse grid regression example with prewavelets from subsection 4.5.2, we encounter two additional problems. First, we cannot simply use the Jackson and Bernstein inequalities (3.42) and (3.40) because of the additional factor k^{m-1} in (3.42). Furthermore, even if we provided valid Jackson and Bernstein inequalities to apply theorem 4.25 and even if we changed the regularization norm to a higher-order Sobolev norm as we discussed above, we would not be able to benefit from $f_\rho \in H^s_{\text{mix}}(T;\mathbb{R}^d)$ for $\frac{3}{2} \leq s \leq 2$ because $V_k \subset \mathcal{H}$ is a necessary condition in theorem 4.25, i.e. the discretization needs to be conforming. However, the prewavelet basis is only contained in $H^s_{\text{mix}}(T;\mathbb{R}^d)$ for $s < \frac{3}{2}$. Therefore, in contrast to our analysis in subsection 4.5.2, we cannot obtain the best possible rate $2^{-2sk}k^{m-1}$ with $\frac{3}{2} \leq s \leq 2$ for the decay of the discretization error by applying theorem 4.25.

4.6 Summary

We conclude this chapter with a brief recapitulation of the methodology and the most important results which we have provided for the constrained regression problem in the preceding sections:

- We introduced the general vector-valued regression problem

$$\text{Find } \hat{f} := \operatorname*{arg\,min}_{f \in L_{2,\rho_T}(T;E)} \mathcal{E}(f) \text{ with } \mathcal{E}(f) := \int_{T\times E} \psi(f(\mathbf{t}),\mathbf{x}) \,\mathrm{d}\rho(\mathbf{t},\mathbf{x}) \tag{A}$$

  and the finite sample problem

$$\text{Find } \operatorname*{arg\,min}_{f \in L_{2,\rho_T}(T;E)} \mathcal{E}_{\mathcal{Z}_n}(f) \text{ with } \mathcal{E}_{\mathcal{Z}_n}(f) := \frac{1}{n}\sum_{i=1}^{n} \psi(f(\mathbf{t}_i),\mathbf{x}_i). \tag{B}$$

  For a qualified cost function $\psi(\mathbf{x},\mathbf{y}) = \tilde{\psi}(\mathbf{x}-\mathbf{y})$ with strictly convex $\tilde{\psi}$, we proved the existence and uniqueness of solutions of (A) if the search set is restricted to a bounded ball $\mathcal{H}_b$ of a real, reflexive Banach space $\mathcal{H}$. In the case of (B), we deduced the existence and uniqueness of a minimal norm solution if $\mathcal{H}$ is a Hilbert space.

- We decomposed the overall regression error into the sum of the bias and the sampling error

$$\underbrace{\mathcal{E}\left(f_{\mathcal{Z}_n,\mathcal{H}_b}\right) - \mathcal{E}(\hat{f})}_{\text{overall error}} = \underbrace{\mathcal{E}(f_{\mathcal{H}_b}) - \mathcal{E}(\hat{f})}_{\text{bias}} + \underbrace{\mathcal{E}(f_{\mathcal{Z}_n,\mathcal{H}_b}) - \mathcal{E}(f_{\mathcal{H}_b})}_{\text{sampling error}}.$$

  Subsequently, we focused on the case $\psi(\mathbf{x},\mathbf{y}) = \|\mathbf{x}-\mathbf{y}\|_E^2$, where the solution of (A) becomes $\hat{f} = f_\rho$. Since the interpolation approach for the bias estimation is not applicable when dealing with finite-dimensional search spaces V_k with $k \in \mathbb{N}$, we considered a novel approach via Jackson and Bernstein estimates. Here, we determined a coupling between the regularization radius b and the discretization parameter k such that the regularization is mild enough to still obtain the optimal (best approximation) rate for the bias/discretization error. Based on the results for the scalar-valued case in [23], we derived estimates for the sampling error of vector-valued regression with fixed confidence $1-\delta$.

- We rigorously analyzed the convergence of constrained regression on sparse grids and hyperbolic crosses. In this generality, the results we obtained are the first of their kind. We analyzed the derived error bounds and observed that the curse of dimensionality only appears with respect to logarithms in n and the basis size N_k. Due to the generality of ρ, we observed that the best possible convergence rate is n^{-1} up to logarithms.

5 Penalized and unregularized regression

In the previous chapter, we tackled the generic regression problem (A) and its finite sample counterpart (B). By restricting the search set to $\mathcal{H}_b$ or $V_{k,b}$, respectively, we obtained a well-posed problem in the sense that a minimizer existed and is unique if one looks for the minimal norm solution. However, we did not yet consider an actual algorithm to solve the regression problem. Since the computational costs of running a constrained, multivariate minimization algorithm are very high, it is usually infeasible to solve (B) over $\mathcal{H}_b$ or $V_{k,b}$ directly. Furthermore, apart from the purely qualitative assertion of well-posedness, we have not made a quantitative statement on the stability of the regression problem yet, which is crucial when considering actual regression algorithms.

Therefore, we introduce the so-called *Lagrangian dual* formulation of the regression problem in this chapter and show how it is related to (B) and how it can be solved. The dual problem is an unconstrained problem by nature, i.e. there is no fixed norm bound incorporated into the search set. Here, regularization is done via an additional penalty term in the minimization functional, which is the reason why it is often also called *penalized* regression. We address the issue of quantitative stability analysis for finite-dimensional search spaces V_k with $k \in \mathbb{N}$ for the penalized problem and the unpenalized problem, where no norm regularization is performed.

One of the main issues when considering a specific regression algorithm, is the choice of the basis of V_k for which the coefficients of the solution are determined. This choice is crucial in the sense that it directly influences the coupling between the number of data points n and the basis size N_k which is needed to obtain stability with high probability. In this context, the works of [19, 21, 60, 61] provide stability and convergence results for specific unregularized regression algorithms if the basis of the search space is chosen to be $L_{2,\rho_T}(T;E)$-orthonormal. In this chapter, we extend the analysis of [21] to obtain stability results also for non-orthonormal bases and for both the penalized and the unpenalized case. Subsequently, we provide an improved convergence rate for unregularized, noiseless function regression, i.e. we assume that the data points $\mathcal{Z}_n = (\mathbf{t}_i, \mathbf{x}_i)_{i=1}^n = (\mathbf{t}_i, g(\mathbf{t}_i))_{i=1}^n \in T \times E$ stem from the evaluation of a function $g : T \to E$. Note that this is a special case of the setting which we had in the last chapter, where $\mathcal{Z}_n$ was drawn according to a general measure ρ. Here, we now assume that $\rho(\mathbf{x}|\mathbf{t}) = \delta_{g(\mathbf{t})}$ is the Dirac measure centered in $g(\mathbf{t})$. The restriction to noiseless function regression leads to faster convergence rates than n^{-1}, which was the limit in the last chapter.

The remainder of this chapter is organized as follows: We introduce the Lagrangian

Table 5.1: Overview on relevant functions, sets and variables for the analysis of the Lagrangian dual problem.

$\mathcal{L}_{\mathcal{Z}_n,b} : \mathcal{H} \times [0,\infty) \to \mathbb{R}$	Lagrangian of (B) over $\mathcal{H}_b$
$\mu \in [0,\infty)$	Lagrange parameter
$f_{\mathcal{Z}_n,V_k,\mu} \in V_k$	minimizer of $\mathcal{L}_{\mathcal{Z}_n,b}(\cdot,\mu)$ over V_k
$\nu_1,\dots,\nu_{N_k} \in V_k$	basis of V_k
$G \in \mathbb{R}^{N_k \times N_k}$	empirical mass matrix of V_k with entries $G_{ij} = \frac{1}{n}\sum_{l=1}^{n} \langle \nu_i(\mathbf{t}_l), \nu_j(\mathbf{t}_l)\rangle_E$
$M \in \mathbb{R}^{N_k \times N_k}$	mass matrix of V_k with entries $M_{ij} = \langle \nu_i, \nu_j \rangle_{L_{2,\rho_T}(T;E)}$
$C \in \mathbb{R}^{N_k \times N_k}$	regularization matrix of V_k with entries $C_{ij} = \langle \nu_i, \nu_j \rangle_{V_k}$
$B : E^n \to \mathbb{R}^{N_k}$	bounded linear operator which fulfills $G = n \cdot B \circ B^*$
$\lambda_{\min}(A), \lambda_{\max}(A) \in \mathbb{R}$	maximum and minimum eigenvalues of a symmetric matrix A
$\kappa(A) \in [1,\infty)$	condition number $\kappa(A) = \frac{\lambda_{\max}(A)}{\lambda_{\min}(A)}$ of a positive definite matrix A
$c_{\frac{1}{2}} \in \mathbb{R}$	constant factor $c_{\frac{1}{2}} = \frac{e^{0.5}}{(1.5)^{1.5}} \approx 0.8975$
$K(N_k) \in (0,\infty)$	basis dependent value $K(N_k) = \sup_{\mathbf{t}\in T} \sum_{i=1}^{N_k} \|\nu_i(\mathbf{t})\|_E^2$
$g : T \to E$	function from which $\mathcal{Z}_n$ is sampled for noiseless regression, i.e. $\mathcal{Z}_n = (\mathbf{t}_i, g(\mathbf{t}_i))_{i=1}^n$
$\tau_r : L_{\infty,\rho_T}(T;E) \to L_{\infty,\rho_T}(T;E)$	truncation operator with threshold r from (4.5)

dual formulation and its relation to the original regression problem (B) in section 5.1. In section 5.2 we describe how the dual problem can be solved over finite-dimensional search spaces V_k. The considerations in section 5.3 are based on the work of [21] and lead to a stability result for function regression with arbitrary bases. Based on these results, we derive an upper bound on the overall regression error for noiseless function regression with finite-dimensional search sets in section 5.4, which complements the results of [21]. Section 5.5 deals with the examination of the examples we introduced in section 4.5. However, this time, we have a look at unregularized noiseless function regression instead of constrained regression. We conclude the chapter with a short summary in section 5.6. A short overview on the new notation which we use in this chapter is given in table 5.1.

5.1 The Lagrangian dual problem

Since the results of this section hold for both infinite- and finite-dimensional search spaces, we use the general notation $\mathcal{H}$ instead of V_k for the search space again. Naturally,

the question arises how to compute the solution $f_{\mathcal{Z}_n,\mathcal{H}_b}$ for a given search space $\mathcal{H}$, a ball radius $b > 0$ and a sample $\mathcal{Z}_n$. Solving the constrained optimization problem (B) over $\mathcal{H}_b$ directly is usually computationally intensive. Therefore, regression algorithms often consider the so-called *dual problem* instead. To this end, we define the Lagrangian $\mathcal{L}_{\mathcal{Z}_n,b} : \mathcal{H} \times [0,\infty) \to \mathbb{R}$ by

$$\mathcal{L}_{\mathcal{Z}_n,b}(f,\mu) := \mathcal{E}_{\mathcal{Z}_n}(f) + \mu\left(\|f\|_{\mathcal{H}}^2 - b^2\right) = \frac{1}{n}\sum_{i=1}^{n}\|f(\mathbf{t}_i) - \mathbf{x}_i\|_E^2 + \mu\left(\|f\|_{\mathcal{H}}^2 - b^2\right). \quad (5.1)$$

For a fixed Lagrange parameter $\mu > 0$, the dual problem now reads

$$\text{Find } \arg\min_{f\in\mathcal{H}} \mathcal{L}_{\mathcal{Z}_n,b}(f,\mu). \quad (5.2)$$

In subsection 5.1.1 we discuss how $f_{\mathcal{Z}_n,\mathcal{H}_b}$, which is usually called the *primal solution* in this context, is related to a solution of (5.2). Subsequently, following the lines of [58], we present the representer theorem for RKHS and explain how a solution of (5.2) can be computed in subsection 5.1.2.

5.1.1 Relation between the primal and the dual problem

To assure that a solution of (5.2) exists and in order to relate it to the primal solution $f_{\mathcal{Z}_n,\mathcal{H}_b}$, we apply the Kuhn-Tucker theorem - theorem 50.A of [90] - to our specific situation. We remind the reader of definition 4.5, where we introduced the concept of sequentially lower semicontinuity.

Theorem 5.1 [Generalized Kuhn-Tucker theorem]
Let A be a closed, convex and non-empty subset of the real, reflexive Banach space X and let $F, G : A \to \mathbb{R}$ be convex and sequentially lower semicontinuous functions. Assume there exists an element $g \in A$ such that $G(g) < 0$. Then the following are equivalent

(i) $\bar{f} \in A$ is a minimizer of

$$\inf_{f\in A, G(f)\le 0} F(f).$$

(ii) There exists a $\bar{\mu} \geq 0$ such that $\bar{f} \in A$ fulfills

$$\mathcal{L}(\bar{f},\bar{\mu}) = \min_{f\in A}\mathcal{L}(f,\bar{\mu}),\ G(\bar{f}) \le 0 \text{ and } \bar{\mu}\cdot G(\bar{f}) = 0,$$

where the Lagrangian is defined by $\mathcal{L}(f,\mu) := F(f) + \mu G(f)$.

Proof. See proposition 50.2 of [90] in combination with proposition 38.7 of [90]. □

Corollary 5.2 [Regression by minimization of the Lagrangian]
Let $\mathcal{H}$ be a real, reflexive Banach space for which the embeddings (4.14) *are continuous*

and let $b > 0$ be fixed. For a qualified cost function $\psi(\mathbf{x}, \mathbf{y}) = \tilde{\psi}(\mathbf{x} - \mathbf{y})$, the following are equivalent:

(i) $\bar{f} \in \mathcal{H}_b$ solves the primal problem (B) *over $\mathcal{H}_b$, i.e. $\bar{f} = f_{\mathcal{Z}_n, \mathcal{H}_b}$.*

(ii) There exists a $\bar{\mu} \geq 0$ such that the dual problem (5.2) *with Lagrange parameter $\mu = \bar{\mu}$ is solved by $\bar{f}$ with $\|\bar{f}\|_{\mathcal{H}} \leq b$ and $\bar{\mu} \cdot \left(\|\bar{f}\|_{\mathcal{H}} - b\right) = 0$.*

Proof. Let $A = X = \mathcal{H}$ and let furthermore

$$F(f) := \mathcal{E}_{\mathcal{Z}_n}(f) = \frac{1}{n} \sum_{i=1}^{n} \tilde{\psi}(f(\mathbf{t}_i) - \mathbf{x}_i)$$

and

$$G(f) := \|f\|_{\mathcal{H}}^2 - b^2.$$

Note that

$$G(f) \leq 0 \;\Leftrightarrow\; \|f\|_{\mathcal{H}}^2 \leq b^2 \;\Leftrightarrow\; \|f\|_{\mathcal{H}} \leq b$$

and also

$$\bar{\mu} \cdot G(f) = 0 \;\Leftrightarrow\; \bar{\mu}(\|f\|_{\mathcal{H}} - b) = 0.$$

Therefore, with our definitions of A, X, F and G the statement of the corollary follows directly from theorem 5.1.

It remains to show that the regression problem meets the prerequisites of theorem 5.1. By definition, A is closed, convex and non-empty because it is a vector space. Furthermore, X is a reflexive Banach space. As we already showed in the proof of corollary 4.7, F is continuous and convex on $\mathcal{H}$. The continuity and convexity of G follow from the corresponding properties of the norm $\|\cdot\|_{\mathcal{H}}$. Finally, note that $g := 0 \in A$ fulfills $G(g) < 0$. Therefore, all prerequisites of theorem 5.1 are met. □

Instead of minimizing (5.2), we can also consider

$$\text{Find} \;\; \arg\min_{f \in \mathcal{H}} \mathcal{E}_{\mathcal{Z}_n}(f) + \mu \|f\|_{\mathcal{H}}^2, \tag{C}$$

as a variant of the dual problem, where we omitted b^2. Note that, for a fixed $\mu > 0$, a minimizer of (C) is also a minimizer of (5.2) and vice versa. It can be shown that the minimizer of (C) is unique if the qualified cost function $\psi(\mathbf{x}, \mathbf{y}) = \tilde{\psi}(\mathbf{x} - \mathbf{y})$ is strictly convex. This can be understood as an analogy to corollary 4.9, where we showed that the minimizer of (B) over $\mathcal{H}_b$ with minimal norm is unique if $\mathcal{H}$ is a Hilbert space.

As mentioned above, considering the dual problem with Lagrange parameter $\mu > 0$ is often preferred over dealing with the primal problem because of the high computational costs of solving the latter. Let us assume that we have computed a minimizer $\bar{f} \in \mathcal{H}$ to the dual problem (C) for $\bar{\mu} > 0$. Then, we observe that $\bar{f}$ fulfills (ii) of corollary 5.2 if

we set $b := \|\bar{f}\|_{\mathcal{H}}$. Therefore, $\bar{f}$ also solves the primal problem (B) over $\mathcal{H}_b$, i.e.

$$f_{\mathcal{Z}_n,\mathcal{H}_{\|\bar{f}\|_{\mathcal{H}}}} = \bar{f}.$$

In conclusion, whenever we have a solution $\bar{f} \in \mathcal{H}$ to the dual problem (C) with a Lagrange parameter $\bar{\mu} > 0$, we know that it also minimizes $\mathcal{E}_{\mathcal{Z}_n}$ over

$$\mathcal{H}_{\|\bar{f}\|_{\mathcal{H}}} = \{f \in \mathcal{H} \mid \|f\|_{\mathcal{H}} \leq \|\bar{f}\|_{\mathcal{H}}\}.$$

In the case where we are given a solution $\bar{f}$ to the dual problem (C) with $\bar{\mu} = 0$, the statement of corollary 5.2 becomes trivial.

5.1.2 The representer theorem

Although we already discussed the Lagrangian dual problem (C) as an alternative to the constrained minimization problem (B) over $\mathcal{H}_b$, the question remains how a solution to (C) can be computed by an actual algorithm. In the special case where $\mathcal{H}$ is a reproducing kernel Hilbert space, the solution of (C) can be computed by the representer theorem, see also [58, 74].

Theorem 5.3 [Representer Theorem]
Let $\mathcal{H}$ be an RKHS which fulfills the prerequisites of proposition 3.18 and let $\mathcal{Z}_n = ((\mathbf{t}_i, \mathbf{x}_i))_{i=1}^n \in (T \times E)^n$ be n samples. Furthermore, let ψ be a qualified cost function. Then, the solution $\bar{f} \in \mathcal{H}$ of (C) *with Lagrange parameter $\bar{\mu} > 0$ is unique and*

$$\bar{f}(\mathbf{t}) = \sum_{i=1}^{n} K(\mathbf{t}_i, \mathbf{t})(\mathbf{c}_i), \tag{5.3}$$

holds for certain coefficients $\mathbf{c}_i \in E, i = 1, \ldots, n$. Here, $K : T \times T \to \mathcal{L}(E, E)$ denotes the kernel of $\mathcal{H}$.

For the special case $\psi(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} - \mathbf{y}\|_E^2$, the coefficients in the representation (5.3) *are given as the solution of the system of linear equations*

$$\sum_{i=1}^{n} \left(K(\mathbf{t}_i, \mathbf{t}_j) + \bar{\mu}\delta_{ij}\right) \mathbf{c}_i = \mathbf{x}_j \quad \forall\, j = 1, \ldots, n, \tag{5.4}$$

where $\delta_{ij} \in \mathcal{L}(E, E)$ is defined as

$$\delta_{ij}(\mathbf{y}) := \begin{cases} \mathbf{y} & \text{if } i = j, \\ \mathbf{0} & \text{else.} \end{cases} \qquad \forall\, \mathbf{y} \in E$$

Proof. The statement of the theorem follows from the more general statements in theorem 4.1 and 4.2 of [58]. □

In the special case of squared norm costs (4.3) and an Euclidean image space $E = \mathbb{R}^d$, theorem 5.3 states that the solution to (C) can be computed by solving a system of linear equations which involves an $n \times n$ block matrix with $d \times d$ matrix entries. Writing down the kernel matrix with scalar entries, we result with an $nd \times nd$ matrix consisting of the blocks $K(\mathbf{t}_i, \mathbf{t}_j) \in \mathbb{R}^{d \times d}$ for every $i, j = 1, \ldots, n$. If, additionally, each component function belongs to a scalar-valued RKHS, each $d \times d$ block is a diagonal matrix, see (3.25).

More generally, the representer theorem 5.3 tells us that the solution to (C), which is a minimization problem over the possibly infinite-dimensional reproducing kernel Hilbert space $\mathcal{H}$, is contained in

$$\operatorname{span}\left\{\operatorname{image}(K_{\mathbf{t}_1}), \ldots, \operatorname{image}(K_{\mathbf{t}_n})\right\},$$

where we interpret $K_{\mathbf{t}_i} : E \to \mathcal{H}$ as a linear functional from E to $\mathcal{H}$, i.e.

$$K_{\mathbf{t}_i}(\mathbf{c}_i) = K(\mathbf{t}_i, \cdot)(\mathbf{c}_i) \in \mathcal{H},$$

see also section 3.3. Therefore, if $\dim(E) < \infty$, the minimization problem (C) becomes a finite-dimensional one - even for an infinite-dimensional search space $\mathcal{H}$.

Although the representer theorem directly explains how the dual problem can be solved in the case of quadratic norm costs (4.3), there are still some drawbacks from the computational point of view: Generally, the kernel matrix is densely populated. Therefore, the runtime for computing the solution of (5.4), e.g. by a QR-algorithm, scales cubically in the number of samples n already in the scalar-valued case $E = \mathbb{R}$. Although there exist sophisticated methods to deal with certain types of kernels, see e.g. [74], the computational complexity is worse than n^2 in general. This makes the kernel approach infeasible if n is large. Furthermore, in some cases a closed formula of the kernel function is not accessible and K is only given by an infinite series expansion. For instance, if $E = \mathbb{R}$, we might have

$$K(\mathbf{s}, \mathbf{t}) = \sum_{l=1}^{\infty} \alpha_l \nu_l(\mathbf{s}) \nu_l(\mathbf{t})$$

for certain functions $\nu_l : T \to \mathbb{R}$ and coefficients $\alpha_l \in \mathbb{R}$, see [43] for several examples. Then, the evaluation of the kernel function at given points involves additional costs and an appropriate truncation of the infinite series expansion has to be done.

The above considerations show that it is reasonable to use another approach than employing the representer theorem in some situations. A more convenient way in such cases is to directly choose finite-dimensional grid-based search spaces V_k for $k \in \mathbb{N}$, e.g. sparse grid spaces as introduced in subsection 3.5.1, and to derive the corresponding system of equations without relying on the kernel representation (5.4). We discuss this case in more detail in the next section.

Summary

In this section, we introduced the dual problem (5.2) and related it to the primal problem (B) over $\mathcal{H}_b$. Furthermore, we considered the representer theorem in order to compute a solution to the dual problem. Let us briefly mention the main results:

- We showed that a solution $\bar{f}$ of the variant (C) of the dual problem (5.2) is also a solution to the primal problem (B) over $\mathcal{H}_b$ with $b = \|\bar{f}\|_{\mathcal{H}}$.
- If $\mathcal{H}$ is a reproducing kernel Hilbert space, (C) can be solved with the help of the representer theorem 5.3, which recasts (C) into a finite-dimensional optimization problem.
- In certain situations, e.g. if the sample size n is large, it is very expensive to compute a solution to the dual problem by solving the system of equations (5.4), which results from the representer theorem for the cost function $\psi(\mathbf{x}, \mathbf{y}) = \|\mathbf{x}-\mathbf{y}\|_E^2$.

5.2 Solving the regression problem in finite-dimensional search spaces

From now on, we again solely consider $\psi(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} - \mathbf{y}\|_E^2$ and we additionally assume that the norm $\|\cdot\|_{V_k}$ of the search space V_k is induced by a corresponding inner product $\langle\cdot,\cdot\rangle_{V_k}$. In the following, we have a look at the system of linear equations which corresponds to the dual regression problem in finite-dimensional search spaces.

5.2.1 The regression problem for arbitrary bases

We consider the dual problem

$$\text{Find } f_{\mathcal{Z}_n,V_k,\mu} := \arg\min_{f\in V_k} \mathcal{E}_{\mathcal{Z}_n}(f)+\mu\|f\|_{V_k}^2 = \arg\min_{f\in V_k} \frac{1}{n}\sum_{i=1}^{n} \|f(\mathbf{t}_i)-\mathbf{x}_i\|_E^2+\mu\|f\|_{V_k}^2, \quad (5.5)$$

i.e. (C) with $\mathcal{H} = V_k$ and $\psi(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} - \mathbf{y}\|_E^2$. We already elaborated on solving this problem with the help of the representer theorem in subsection 5.1.2. Now, we use the fact that the V_k are finite-dimensional to obtain a different system of equations instead. To this end, let $k \in \mathbb{N}$ be fixed and let $\nu_j^{(k)}, j = 1, \ldots, N_k$ be a basis of V_k. For the ease of notation, we will omit the upper index (k) and write ν_j in the following.

Proposition 5.4 [System of linear equations for the dual problem]
The solution $f_{\mathcal{Z}_n,V_k,\mu} = \sum_{j=1}^{N_k} \alpha_j\nu_j$ to (5.5) can be computed by solving the system of linear equations

$$(G + \mu C)\vec{\alpha} = \vec{\nu}_{\mathbf{x}}, \quad (5.6)$$

where $G, C \in \mathbb{R}^{N_k \times N_k}$ *and* $\vec{\alpha}, \vec{\nu}_{\mathbf{x}} \in \mathbb{R}^{N_k}$ *are given by*[1]

$$G_{ij} = \frac{1}{n}\sum_{l=1}^{n}\langle \nu_i(\mathbf{t}_l), \nu_j(\mathbf{t}_l)\rangle_E, \quad C_{ij} = \langle \nu_i, \nu_j\rangle_{V_k}, \quad \vec{\alpha}_i = \alpha_i, \quad (\vec{\nu}_{\mathbf{x}})_i = \frac{1}{n}\sum_{l=1}^{n}\langle \nu_i(\mathbf{t}_l), \mathbf{x}_l\rangle_E$$

for all $i, j = 1, \ldots, N_k$.

Proof. The proof works analogously to the scalar-valued case $E = \mathbb{R}$, which can be found in e.g. [35]. To this end, note that a minimizer of (5.5) fulfills

$$\begin{aligned}
0 &= \frac{\partial}{\partial \alpha_i}\left(\frac{1}{n}\sum_{l=1}^{n}\left\|\sum_{p=1}^{N_k}\alpha_p\nu_p(\mathbf{t}_l) - \mathbf{x}_l\right\|_E^2 + \mu\left\|\sum_{p=1}^{N_k}\alpha_p\nu_p\right\|_{V_k}^2\right) \\
&= \frac{\partial}{\partial \alpha_i}\left(\frac{1}{n}\sum_{l=1}^{n}\left(\sum_{p,q=1}^{N_k}\alpha_p\alpha_q\langle\nu_p(\mathbf{t}_l), \nu_q(\mathbf{t}_l)\rangle_E - 2\sum_{p=1}^{N_k}\alpha_p\langle\nu_p(\mathbf{t}_l), \mathbf{x}_l\rangle_E + \langle\mathbf{x}_l, \mathbf{x}_l\rangle_E\right)\right) \\
&\quad + \frac{\partial}{\partial \alpha_i}\left(\mu\sum_{p,q=1}^{N_k}\alpha_p\alpha_q\langle\nu_p, \nu_q\rangle_{V_k}\right) \\
&= \frac{1}{n}\sum_{l=1}^{n}\left(2\sum_{p=1}^{N_k}\alpha_p\langle\nu_p(\mathbf{t}_l), \nu_i(\mathbf{t}_l)\rangle_E - 2\langle\nu_i(\mathbf{t}_l), \mathbf{x}_l\rangle_E\right) + \mu\sum_{p=1}^{N_k}2\alpha_p\langle\nu_p, \nu_i\rangle_{V_k}
\end{aligned}$$

for all $i = 1, \ldots, N_k$. This proves the assertion. □

After assembling the vectors and matrices in (5.6), the coefficients of $f_{\mathcal{Z}_n, V_k, \mu}$ can be computed by a direct (e.g. via QR-decomposition) or an iterative (e.g. via conjugate gradients) system solver. The main difference in contrast to the system (5.4), which results from the representer theorem, is that the assemblation of all matrices scales only linearly in the number of sample points n. The choice of an appropriate solver depends on the choice of the basis $\nu_1, \ldots, \nu_{N_k}$ in proposition 5.4. For localized basis functions, such as the piecewise linear splines, the matrices in (5.6) are usually sparsely populated and an iterative solver is able to compute $f_{\mathcal{Z}_n, V_k, \mu}$ with significantly lower computational costs than a direct method, which scales like N_k^3, see [70] for details.

5.2.2 Operator splitting

In the following, we will interpret the samples $\vec{\mathbf{x}} := (\mathbf{x}_1, \ldots, \mathbf{x}_{N_k})^T$ as an element of the product space $E^n = E \times \ldots \times E$. To this end, let

$$\langle \vec{\mathbf{a}}, \vec{\mathbf{b}}\rangle_{E^n} := \sum_{i=1}^{n}\langle \mathbf{a}_i, \mathbf{b}_i\rangle_E,$$

[1] This result has to be seen in analogy to (5.4): In the case $E = \mathbb{R}$ for example, the basis functions can be chosen as $\nu_j(\cdot) = K(\mathbf{t}_j, \cdot)$, where K is the kernel of V_k.

such that E^n also becomes a Hilbert space with this specific inner product.

Proposition 5.5
For a fixed sample $\mathcal{Z}_n$, let the linear operator $B : E^n \to \mathbb{R}^{N_k}$ be given by

$$B(\vec{\mathbf{a}}) := \frac{1}{n} \sum_{i=1}^{n} \begin{pmatrix} \langle \nu_1(\mathbf{t}_i), \mathbf{a}_i \rangle_E \\ \vdots \\ \langle \nu_{N_k}(\mathbf{t}_i), \mathbf{a}_i \rangle_E \end{pmatrix}. \tag{5.7}$$

Then, B is bounded and the adjoint operator $B^* : \mathbb{R}^{N_k} \to E^n$ is given by

$$B^*(\vec{\beta}) := \frac{1}{n} \sum_{j=1}^{N_k} \beta_j \begin{pmatrix} \nu_j(\mathbf{t}_1) \\ \vdots \\ \nu_j(\mathbf{t}_n) \end{pmatrix}. \tag{5.8}$$

Furthermore, it holds $G = n \cdot B \circ B^*$ and $\vec{\nu}_{\mathbf{x}} = B(\vec{\mathbf{x}})$ in proposition 5.4 with $\vec{\mathbf{x}} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)^T$. Therefore, (5.6) can be rewritten as

$$(n \cdot B \circ B^* + \mu C)(\vec{\alpha}) = B(\vec{\mathbf{x}}). \tag{5.9}$$

Proof. Let $\vec{\mathbf{a}} \in E^n$. Note that the triangle inequality and the Cauchy-Schwarz inequality lead to

$$\begin{aligned} \|B(\vec{\mathbf{a}})\|_{\mathbb{R}^{N_k}} &= \left\| \frac{1}{n} \sum_{i=1}^{n} \begin{pmatrix} \langle \nu_1(\mathbf{t}_i), \mathbf{a}_i \rangle_E \\ \vdots \\ \langle \nu_{N_k}(\mathbf{t}_i), \mathbf{a}_i \rangle_E \end{pmatrix} \right\|_{\mathbb{R}^{N_k}} \leq \frac{1}{n} \sum_{i=1}^{n} \left\| \begin{pmatrix} \langle \nu_1(\mathbf{t}_i), \mathbf{a}_i \rangle_E \\ \vdots \\ \langle \nu_{N_k}(\mathbf{t}_i), \mathbf{a}_i \rangle_E \end{pmatrix} \right\|_{\mathbb{R}^{N_k}} \\ &= \frac{1}{n} \sum_{i=1}^{n} \sqrt{\sum_{j=1}^{N_k} \langle \nu_j(\mathbf{t}_i), \mathbf{a}_i \rangle_E^2} \leq \frac{1}{n} \sum_{i=1}^{n} \sqrt{\sum_{j=1}^{N_k} \|\nu_j(\mathbf{t}_i)\|_E^2 \|\mathbf{a}_i\|_E^2} \\ &\leq \frac{1}{n} \left(\max_{i=1,\ldots,n} \sqrt{\sum_{j=1}^{N_k} \|\nu_j(\mathbf{t}_i)\|_E^2} \right) \cdot \sum_{i=1}^{n} \|\mathbf{a}_i\|_E \leq c \cdot \|\vec{\mathbf{a}}\|_{E^n} \end{aligned}$$

with a prefactor $c := \frac{1}{\sqrt{n}} \max_{i=1,\ldots,n} \sqrt{\sum_{j=1}^{N_k} \|\nu_j(\mathbf{t}_i)\|_E^2}$. Here, the $\frac{1}{\sqrt{n}}$ in c appears because an additional factor of $\sqrt{n}$ has to be introduced when bounding the ℓ_1-type norm $\sum_{i=1}^{n} \|\mathbf{a}_i\|_E$ by the ℓ_2-type norm $\|\vec{\mathbf{a}}\|_{E^n}$. This proves the boundedness of B since c is a constant for a fixed sample $\mathcal{Z}_n$ and a fixed search space V_k. Now, let $\vec{\beta} \in \mathbb{R}^{N_k}$ be arbitrary. Since B is a bounded, linear operator between Hilbert spaces, the adjoint operator $B^* : \mathbb{R}^{N_k} \to E^n$ exists and fulfills

$$\langle \vec{\mathbf{a}}, B^*(\vec{\beta}) \rangle_{E^n} = \langle B(\vec{\mathbf{a}}), \vec{\beta} \rangle_{\mathbb{R}^{N_k}} = \frac{1}{n} \sum_{j=1}^{N_k} \sum_{i=1}^{n} \langle \nu_j(\mathbf{t}_i), \mathbf{a}_i \rangle_E \cdot \beta_j.$$

Because of the definition of $\langle \cdot, \cdot \rangle_{E^n}$, (5.8) follows immediately. Finally, note that $\vec{\nu}_{\mathbf{x}} = B(\vec{\mathbf{x}})$ follows directly from the definition of B and we also have

$$\begin{aligned} B \circ B^*(\vec{\beta}) &= B\left(\frac{1}{n}\sum_{j=1}^{N_k}\beta_j \begin{pmatrix} \nu_j(\mathbf{t}_1) \\ \vdots \\ \nu_j(\mathbf{t}_n) \end{pmatrix}\right) = \frac{1}{n}\sum_{i=1}^{n} \begin{pmatrix} \langle \nu_1(\mathbf{t}_i), \frac{1}{n}\sum_{j=1}^{N_k}\beta_j\nu_j(\mathbf{t}_i)\rangle_E \\ \vdots \\ \langle \nu_{N_k}(\mathbf{t}_i), \frac{1}{n}\sum_{j=1}^{N_k}\beta_j\nu_j(\mathbf{t}_i)\rangle_E \end{pmatrix} \\ &= \frac{1}{n^2}\sum_{j=1}^{N_k}\beta_j \sum_{i=1}^{n} \begin{pmatrix} \langle \nu_1(\mathbf{t}_i), \nu_j(\mathbf{t}_i)\rangle_E \\ \vdots \\ \langle \nu_{N_k}(\mathbf{t}_i), \nu_j(\mathbf{t}_i)\rangle_E \end{pmatrix} = \frac{1}{n} G \cdot \vec{\beta}, \end{aligned}$$

which proves $G = n \cdot B \circ B^*$. □

Summary

The main results of this section can be summarized as follows:

- We provided an alternate approach to the representer theorem in order to recast the dual regression problem (C) into a system of linear equations if the search space is finite-dimensional. The corresponding $N_k \times N_k$ system is given by (5.6). The computational costs for assembling and solving this system scale only linearly in n.

- With the help of the linear operators B and B^* from (5.7) and (5.8), we deduced the different notation (5.9) for the linear equation system (5.6).

5.3 Stability analysis

With the help of the results from the previous section, we now analyze the stability of the regression problem in finite-dimensional search spaces. To this end, we build our analysis on a result of [83], which provides a probability bound for the deviation from the mean for eigenvalues of random matrices. Here, we mainly follow the lines of [21], which first used the bounds of [83] to establish stability and convergence results for scalar-valued, unregularized regression, i.e. $\mu = 0$ and $E = \mathbb{R}$. The analysis in [21] is restricted to an $L_{2,\rho_T(T)}$-orthonormal basis $\nu_1, \ldots, \nu_{N_k}$. We extend their results to the vector-valued case and obtain estimates also for regularized regression with arbitrary basis.

5.3.1 A Chernoff inequality for random matrices

We recapitulate the matrix Chernoff bound given in [83]. To this end, we denote by $\lambda_{\max}(X)$ and $\lambda_{\min}(X)$ the maximum and minimum eigenvalue of a symmetric matrix X.

Theorem 5.6 [Chernoff inequality for random matrices]
Let $D \in \mathbb{N}$ and let $X_1, \dots, X_n \in \mathbb{R}^{D\times D}$ be a collection of independent, symmetric, positive semidefinite matrices with random entries. Let, furthermore, $R > 0$ be such that $\lambda_{\max}(X_i) \leq R$ for all $i = 1, \dots, n$. Then, for $\delta \in [0, 1)$, it holds

$$\mathbb{P}\left[\lambda_{\min}\left(\sum_{i=1}^{n} X_i\right) \leq (1-\delta)c_{\min}\right] \leq D\left(\frac{e^{-\delta}}{(1-\delta)^{1-\delta}}\right)^{\frac{c_{\min}}{R}}$$

and

$$\mathbb{P}\left[\lambda_{\max}\left(\sum_{i=1}^{n} X_i\right) \geq (1+\delta)c_{\max}\right] \leq D\left(\frac{e^{\delta}}{(1+\delta)^{1+\delta}}\right)^{\frac{c_{\max}}{R}}$$

with $c_{\min} := \lambda_{\min}\left(\mathbb{E}\left[\sum_{i=1}^{n} X_i\right]\right)$ and $c_{\max} := \lambda_{\max}\left(\mathbb{E}\left[\sum_{i=1}^{n} X_i\right]\right)$.

Proof. The proof can be found in section 5 of [83]. □

In the following, we consider the spectral norm for matrices induced by the Euclidean vector norm, i.e. we write

$$\|A\|_2 := \|A\|_{\mathcal{L}\left(\mathbb{R}^{N_k}, \mathbb{R}^{N_k}\right)} \overset{(3.23)}{=} \max_{\|\mathbf{x}\|_2 = 1} \|A\mathbf{x}\|_2 = \sqrt{\lambda_{\max}(A^T A)},$$

which is equal to $\lambda_{\max}(A)$ for symmetric, positive semidefinite A. Therefore, we additionally have $\|A^{-1}\|_2 = \lambda_{\min}(A)^{-1}$ if A is positive definite.

Corollary 5.7
With the notation of proposition 5.4, let $X, M \in \mathbb{R}^{N_k \times N_k}$ be defined by

$$X_{ij} := \frac{1}{n}\langle \nu_i(\mathbf{t}), \nu_j(\mathbf{t})\rangle_E \quad \forall\, i, j = 1, \dots, N_k,$$

where $\mathbf{t}$ is drawn according to ρ_T and let

$$M_{ij} := \langle \nu_i, \nu_j \rangle_{L_{2,\rho_T}(T;E)} \quad \forall\, i, j = 1, \dots, N_k.$$

If $R > 0$ fulfills $R \geq \lambda_{\max}(X)$ almost surely, we obtain

$$P := \mathbb{P}\left[\|G\|_2 \geq \frac{3}{2}\lambda_{\max}(M) \quad or \quad \|G^{-1}\|_2 \geq \frac{2}{\lambda_{\min}(M)}\right] \leq 2N_k c_{\frac{1}{2}}^{\frac{\lambda_{\min}(M)}{R}}$$

with $c_{\frac{1}{2}} := \frac{e^{0.5}}{(1.5)^{1.5}} \approx 0.8975$.

Proof. Note that X is positive semidefinite since

$$\vec{\beta}^T X \vec{\beta} = \frac{1}{n}\left\langle \sum_{i=1}^{N_k} \beta_i \nu_i(\mathbf{t}), \sum_{i=1}^{N_k} \beta_i \nu_i(\mathbf{t}) \right\rangle_E = \frac{1}{n}\left\| \sum_{i=1}^{N_k} \beta_i \nu_i(\mathbf{t}) \right\|_E^2 \geq 0$$

for all $\vec{\beta} \in \mathbb{R}^{N_k}$. Note, furthermore, that the so-called mass matrix M is positive definite because we have

$$\vec{\beta}^T M \vec{\beta} = \left\langle \sum_{i=1}^{N_k} \beta_i \nu_i, \sum_{i=1}^{N_k} \beta_i \nu_i \right\rangle_{L_{2,\rho_T}(T;E)} = \left\| \sum_{i=1}^{N_k} \beta_i \nu_i \right\|^2_{L_{2,\rho_T}(T;E)} \geq 0,$$

which is 0 if and only if $\vec{\beta} = \vec{0}$ since the ν_i are linearly independent. We now take $X_1, \ldots, X_n$ to be independent drawings of the random matrix X. Then, with $D = N_k$, the prerequisites of theorem 5.6 are fulfilled. Note that $G = \sum_{i=1}^n X_i$ if the samples are taken to be $\mathcal{Z}_n$. Furthermore, we have $M = \mathbb{E}\left[\sum_{i=1}^n X_i\right]$. Since M is positive definite, we obtain

$$\lambda_{\min}(G) > \frac{1}{2}\lambda_{\min}(M) \Leftrightarrow G^{-1} \text{ exists and } \|G^{-1}\|_2 = \frac{1}{\lambda_{\min}(G)} < \frac{2}{\lambda_{\min}(M)}.$$

Therefore, choosing $\delta = \frac{1}{2}$, we obtain

$$P \leq N_k \left(\frac{e^{0.5}}{(1.5)^{1.5}} \right)^{\frac{\lambda_{\max}(M)}{R}} + N_k \left(\frac{e^{-0.5}}{(0.5)^{0.5}} \right)^{\frac{\lambda_{\min}(M)}{R}}$$

from theorem 5.6. Since $1 > c_{\frac{1}{2}} = \frac{e^{0.5}}{(1.5)^{1.5}} > \frac{e^{-0.5}}{(0.5)^{0.5}} \approx 0.858 > 0$ and since $\lambda_{\max}(M) \geq \lambda_{\min}(M)$, we finally obtain

$$P \leq N_k c_{\frac{1}{2}}^{\frac{\lambda_{\max}(M)}{R}} + N_k c_{\frac{1}{2}}^{\frac{\lambda_{\min}(M)}{R}} \leq 2N_k c_{\frac{1}{2}}^{\frac{\lambda_{\min}(M)}{R}},$$

which concludes the proof. □

Corollary 5.7 tells us that the spectral norms of both G and G^{-1} are bounded with high probability. However, we still need an adequate bound R for the maximum eigenvalue of the random matrix X. To this end, let us define a characteristic number, which will be useful in the following.

Definition 5.8 [The bound $K(N_k)$]
Let the basis $\nu_1, \ldots, \nu_{N_k}$ of V_k be fixed. We define the number $K(N_k)$ by

$$K(N_k) := \sup_{\mathbf{t} \in T} \sum_{i=1}^{N_k} \|\nu_i(\mathbf{t})\|_E^2. \tag{5.10}$$

Now, we relate $K(N_k)$ to R from corollary 5.7. To this end, we decompose X into two linear operators. This is similar to the decomposition of G into B^* and B from proposition 5.5.

Lemma 5.9
Let $X \in \mathbb{R}^{N_k \times N_k}$ be defined by

$$X_{ij} := \frac{1}{n}\langle \nu_i(\mathbf{t}), \nu_j(\mathbf{t})\rangle_E.$$

as in corollary 5.7. Then, $\lambda_{\max}(X) = \|X\|_2 \leq \frac{1}{n}K(N_k)$.

Proof. Let $\mathbf{t} \in T$ be randomly drawn according to ρ_T. Let $A : E \to \mathbb{R}^{N_k}$ be the bounded[2], linear map defined by

$$A(\mathbf{x}) := \frac{1}{n}\begin{pmatrix} \langle \nu_1(\mathbf{t}), \mathbf{x}\rangle_E \\ \vdots \\ \langle \nu_{N_k}(\mathbf{t}), \mathbf{x}\rangle_E \end{pmatrix}^T.$$

Since $\frac{1}{n}\sum_{i=1}^{N_k} \beta_i \langle \nu_i(\mathbf{t}), \mathbf{x}\rangle_E = \langle A(\mathbf{x}), \vec{\beta}\rangle_{\mathbb{R}^{N_k}} = \langle \mathbf{x}, A^*(\vec{\beta})\rangle_E$ for all $\vec{\beta} \in \mathbb{R}^{N_k}$ and all $\mathbf{x} \in E$, the adjoint operator $A^* : \mathbb{R}^{N_k} \to E$ is given by

$$A^*(\vec{\beta}) := \frac{1}{n}\sum_{i=1}^{N_k} \beta_i \nu_i(\mathbf{t}).$$

Furthermore, we have

$$A \circ A^*(\vec{\beta}) = A\left(\frac{1}{n}\sum_{i=1}^{N_k} \beta_i \nu_i(\mathbf{t})\right) = \frac{1}{n^2}\sum_{i=1}^{N_k} \beta_i \begin{pmatrix} \langle \nu_1(\mathbf{t}), \nu_i(\mathbf{t})\rangle_E \\ \vdots \\ \langle \nu_{N_k}(\mathbf{t}), \nu_i(\mathbf{t})\rangle_E \end{pmatrix} = \frac{1}{n}X\vec{\beta}$$

for all $\vec{\beta} \in \mathbb{R}^{N_k}$. Therefore, we can estimate the spectral norm of X by

$$\begin{aligned} \|X\|_2 = \|nA \circ A^*\|_2 = n\|A\|^2_{\mathcal{L}(E,\mathbb{R}^{N_k})} &= n \max_{\|\mathbf{x}\|_E=1} \|A(\mathbf{x})\|_2^2 \\ = \frac{1}{n}\max_{\|\mathbf{x}\|_E=1} \sum_{i=1}^{N_k} \langle \nu_i(\mathbf{t}), \mathbf{x}\rangle_E^2 &\leq \frac{1}{n}\max_{\|\mathbf{x}\|_E=1} \sum_{i=1}^{N_k} \|\nu_i(\mathbf{t})\|_E^2 \|\mathbf{x}\|_E^2 \leq \frac{1}{n}K(N_k), \end{aligned}$$

which proves the assertion. □

Lemma 5.9 states that $R = \frac{1}{n}K(N_k)$ is a valid choice in corollary 5.7. Note that the size $K(N_k)$ generally depends on the choice of the basis $\nu_1, \ldots, \nu_{N_k}$ in contrast to the special case, where only orthonormal bases are allowed, see [21] for details. To apply corollary 5.7, we need to fix a basis and estimate $K(N_k)$. We conclude this subsection with a reformulation of the Chernoff matrix bound in terms of a condition on $K(N_k)$.

[2]The fact that A is bounded can be seen easily by analogous arguments as in the proof of proposition 5.5.

Lemma 5.10
Let the prerequisites of corollary 5.7 be fulfilled, let $n \geq N_k$ and let $\sigma > 0$ be such that

$$K(N_k) \leq \frac{\lambda_{\min}(M)|\log(c_{\frac{1}{2}})|}{(1+\sigma)} \cdot \frac{n}{\log(n)}. \tag{5.11}$$

Then

$$P := \mathbb{P}\left[\|G\|_2 \geq \frac{3}{2}\lambda_{\max}(M) \quad or \quad \|G^{-1}\|_2 \geq \frac{2}{\lambda_{\min}(M)}\right] \leq 2n^{-\sigma}.$$

Proof. Using the fact that $\log(c_{\frac{1}{2}}) < 0$ and combining corollary 5.7 and lemma 5.9, we obtain with $R := \frac{K(N_k)}{n}$

$$\begin{aligned} P &\leq 2N_k c_{\frac{1}{2}}^{\frac{n\lambda_{\min}(M)}{K(N_k)}} = 2N_k \exp\left(\frac{\log(c_{\frac{1}{2}})n\lambda_{\min}(M)}{K(N_k)}\right) \\ &\leq 2N_k \exp\left(-(1+\sigma)\log(n)\right) = 2N_k n^{-(1+\sigma)} \leq 2n^{-\sigma} \end{aligned}$$

since we assumed $N_k \leq n$. □

Note that $n \geq N_k$ is a necessary prerequisite in lemma 5.10. This can be seen in analogy to the necessary conditions for convergence in section 4.5. Note that n is coupled to N_k also via (5.11). As we will see, this automatically implies the restriction $n \geq N_k$ for our examples.

5.3.2 Stability of the regression problem

We now combine the results of the previous two subsections to obtain a statement on the stability of the least-squares regression problem in finite-dimensional search spaces.

Theorem 5.11 [Stability of least-squares regression]
Let $n \geq N_k$ and let $K(N_k)$ fulfill (5.11). Then, the solution $f_{\mathcal{Z}_n,V_k,\mu} = \sum_{j=1}^{N_k} \alpha_j \nu_j$ of (5.6) fulfills

$$\|f_{\mathcal{Z}_n,V_k,\mu}\|_{L_{2,\rho_T}(T;E)} \leq \frac{\sqrt{6}\lambda_{\max}(M)}{\lambda_{\min}(M) + \mu \cdot \lambda_{\min}(C)} \cdot \frac{1}{\sqrt{n}}\|\vec{\mathbf{x}}\|_{E^n}$$

with probability at least $1 - 2n^{-\sigma}$, where we denote $\vec{\mathbf{x}} := (\mathbf{x}_1, \ldots, \mathbf{x}_n)^T$.

Proof. By lemma 5.10, we have

$$\mathbb{P}\left[\|G\|_2 < \frac{3}{2}\lambda_{\max}(M) \quad \text{and} \quad \|G^{-1}\|_2 < \frac{2}{\lambda_{\min}(M)}\right] \geq 1 - 2n^{-\sigma}.$$

Let $S := n \cdot B \circ B^* + \mu C \in \mathbb{R}^{N_k \times N_k}$ be the system matrix. Since $G = n \cdot B \circ B^*$, see proposition 5.5, and since M resembles the mass matrix for the basis $\nu_1, \ldots, \nu_{N_k}$, we

obtain

$$\begin{aligned}\|f_{\mathcal{Z}_n,V_k,\mu}\|^2_{L_{2,\rho_T}(T;E)} &= \vec{\alpha}^T M \vec{\alpha} \overset{(5.9)}{=} B(\vec{\mathbf{x}})^T S^{-1} M S^{-1} B(\vec{\mathbf{x}}) \\ &\leq \|B\|^2_{\mathcal{L}\left(E^n,\mathbb{R}^{N_k}\right)} \|S^{-1}\|_2^2 \|M\|_2 \|\vec{\mathbf{x}}\|^2_{E^n} \\ &= \frac{1}{n}\|G\|_2 \|S^{-1}\|_2^2 \|M\|_2 \|\vec{\mathbf{x}}\|^2_{E^n} \\ &\leq \frac{3}{2n}\lambda_{\max}(M)^2 \|S^{-1}\|_2^2 \|\vec{\mathbf{x}}\|^2_{E^n}\end{aligned}$$

with probability greater or equal to $1-2n^{-\sigma}$. Therefore, it suffices to show that $\|S^{-1}\|_2 \leq \frac{2}{\lambda_{\min}(M)+\mu\cdot\lambda_{\min}(C)}$ to complete the proof. To this end, note that G is positive definite with probability at least $1-2n^{-\sigma}$ since M is positive definite. Furthermore, C is positive definite by definition. Therefore, $S = G + \mu C$ is invertible and we obtain

$$\begin{aligned}\frac{1}{\|S^{-1}\|_2} = \lambda_{\min}(S) &\geq \lambda_{\min}(G) + \mu\cdot\lambda_{\min}(C) \geq \frac{1}{2}\lambda_{\min}(M) + \mu\cdot\lambda_{\min}(C) \\ &\geq \frac{\lambda_{\min}(M) + \mu\cdot\lambda_{\min}(C)}{2},\end{aligned}$$

which concludes the proof. □

Theorem 5.11 gives us an upper bound on the $L_{2,\rho_T}(T)$ norm of the solution of the regularized regression problem (C) in terms of a prefactor and the E^n norm of the input data $\vec{\mathbf{x}}$. Since $\frac{1}{\sqrt{n}}\|\vec{\mathbf{x}}\|_{E^n} = \sqrt{\frac{1}{n}\sum_{i=1}^n \|\mathbf{x}_i\|_E^2} \leq r$ almost surely with r from (4.5), the regularized regression problem is stable with high probability if the factor

$$\frac{\lambda_{\max}(M)}{\lambda_{\min}(M) + \mu\cdot\lambda_{\min}(C)}$$

is small or at least bounded from above for $k \to \infty$. There are two different ways to ensure this:

1. The mass matrix M is such that its condition number $\kappa(M) := \frac{\lambda_{\max}(M)}{\lambda_{\min}(M)}$ is bounded from above for the whole scale $(V_k)_{k=1}^\infty$ of search spaces. In this case, the problem is stable with high probability also without regularization ($\mu = 0$). Recall that $M_{ij} = \langle \nu_i, \nu_j \rangle_{L_{2,\rho_T}(T;E)}$. Therefore, we have $\kappa(M) = 1$ for orthonormal bases, e.g. for the Fourier basis, which we considered in subsection 3.5.2. This case is essentially covered also by the analysis of [21]. However, the mass matrix condition number is bounded from above also for Riesz bases, e.g. for the prewavelet basis, which we introduced in subsection 3.5.1. The Riesz property can be written as

$$\|\vec{\alpha}\|_2^2 \simeq \left\| \sum_{i=1}^{N_k} \alpha_i \nu_i \right\|^2_{L_{2,\rho_T}(T;E)}$$

for all V_k with $k \in \mathbb{N}$, where the equivalence constants implied by $\simeq$ do not depend on k. Since we have

$$\left\| \sum_{i=1}^{N_k} \alpha_i \nu_i \right\|^2_{L_{2,\rho_T}(T;E)} = \vec{\alpha}^T M \vec{\alpha},$$

these equivalence constants are exactly $\lambda_{\max}(M)$ and $\lambda_{\min}(M)$. Therefore, also Riesz bases can be employed in theorem 5.11 to obtain a stable regression problem.

2. The regularization is chosen such that $\mu \cdot \lambda_{\min}(C)$ compensates the degeneration of the condition number $\kappa(M)$ of the mass matrix. An example, where the mass matrix condition becomes worse with $k \to \infty$ but the regularization matrix stabilizes this effect, is the case of H^1 regularized regression with hat functions. To this end, we take $m = 1, d = 1, T = (0, 1)$ and consider the space of piecewise linear splines $\mathcal{V}_k^{\text{full}}$. However, instead of taking the hierarchical prewavelet basis $\gamma_{l,i}$ with $|l|_{\ell_1} \leq k, i \in I_l$, which we introduced in subsection 3.5.1, we take the hierarchical hat function basis, i.e. we substitute $\gamma_{l,i}$ by $\phi_{l,i}$, see (3.28). Since the unscaled hat functions do not form a Riesz basis of V_k, the condition number of the mass matrix is not bounded from above uniformly in $k \in \mathbb{N}$. To see this, let $\rho_T = \lambda_T$ be the Lebesgue measure. Since the diagonal entries of the mass matrix are

$$\int_T \phi_{l,i}^2(\mathbf{t}) \, \mathrm{d}\rho_T(\mathbf{t}) = \frac{1}{3} \text{ if } l = 0 \quad \text{and} \quad \int_T \phi_{l,i}^2(\mathbf{t}) \, \mathrm{d}\rho_T(\mathbf{t}) = \frac{2}{3} \cdot 2^{-l} \text{ if } l > 0,$$

we obtain that $\lambda_{\max}(M) \geq \frac{1}{3}$ and $\lambda_{\min}(M) \leq \frac{2}{3} \cdot 2^{-k}$. Thus, we have $2^k \lesssim \kappa(M)$ and the mass matrix condition number deteriorates for $k \to \infty$. However, the regularization matrix C, which corresponds to the H^1 norm, is a sum of M, which accounts for the L_2 part of the H^1 norm, and an almost[3] diagonal scaling matrix with entries 2^{l+1} for the basis functions on level $l \in \{0, \ldots, k\}$. The latter resembles the derivative term of the H^1 norm. It can easily be seen that the smallest eigenvalue $\lambda_{\min}(C)$ is bounded from below independently of $k \in \mathbb{N}$. Thus,

$$\frac{\lambda_{\max}(M)}{\lambda_{\min}(M) + \mu \cdot \lambda_{\min}(C)}$$

can be controlled by the size of the regularization parameter $\mu > 0$. For more details on this specific regularization and variants thereof as well as the corresponding numerical treatment, we refer the interested reader to [35, 36, 68].

Summary

In this section, we investigated the stability of the penalized regression problem (C) in dependence on the coupling between n and N_k, the choice of the basis $\nu_1, \ldots, \nu_{N_k}$ of the

[3] The only non-diagonal non-zero entries are the ones for the basis functions on level 0.

search space V_k and the choice of the regularization parameter $\mu > 0$. We summarize the most important results.

- We used the matrix Chernoff inequality derived in [83] to get probabilistic bounds on the condition number of the empirical mass matrix G, which converges to the mass matrix M when the number of samples n tends to ∞.

- Based on the techniques of [21], we derived a stability result for solving the system of equations corresponding to the dual least-squares regression problem (C). We extended the results of [21] in the sense that theorem 5.11 can also be applied in the vector-valued case and also with non-$L_{2,\rho_T}(T;E)$-orthonormal bases for both regularized and unregularized regression.

- If the size of $K(N_k)$ can be bounded by (5.11), the least-squares regression problem is stable with high probability if we are dealing with a Riesz basis $\nu_1, \dots, \nu_{N_k}$ or if the regularization parameter μ is chosen large enough.

5.4 Noiseless function regression

In order to derive convergence rates which exceed the ones from chapter 4, we now consider the case of noiseless function regression over finite-dimensional search spaces V_k for $k \in \mathbb{N}$. Here, we restrict ourselves to the unregularized case, i.e. $\mu = 0$. For noiseless function regression, there exists a point-evaluable $g \in L_{\infty,\rho_T}(T;E)$ with $\mathbf{x}_i = g(\mathbf{t}_i)$ and, thus, $\mathcal{Z}_n = (\mathbf{t}_i, g(\mathbf{t}_i))_{i=1}^n$. Then, we directly obtain $\rho(\mathbf{x}|\mathbf{t}) = \delta_{g(\mathbf{t})}(\mathbf{x})$ and $f_\rho = g$. Therefore, the regression problem (C) becomes

$$\text{Find} \quad f_{\mathcal{Z}_n,V_k} = \operatorname*{arg\,min}_{f\in V_k} \mathcal{E}_{\mathcal{Z}_n}(f) = \operatorname*{arg\,min}_{f\in V_k} \frac{1}{n}\sum_{i=1}^{n} \|f(\mathbf{t}_i) - g(\mathbf{t}_i)\|_E^2. \tag{D}$$

Note that this is exactly (B) with search set V_k, $\psi(\mathbf{x},\mathbf{y}) = \|\mathbf{x}-\mathbf{y}\|_E^2$ and $\mathbf{x}_i = g(\mathbf{t}_i)$. Since $\mu = 0$, the corresponding system of linear equations can be written as

$$G\vec{\alpha} = B\left(\overrightarrow{\boldsymbol{g}(\mathbf{t})}\right) \tag{5.12}$$

with $\overrightarrow{\boldsymbol{g}(\mathbf{t})} = \vec{\mathbf{x}} = (g(\mathbf{t}_1), \dots, g(\mathbf{t}_n))^T$, cf. (5.6), (5.9). The solution to the regression problem can then be written as $f_{\mathcal{Z}_n,V_k} = \sum_{i=1}^{N_k} \alpha_i \nu_i$. Note, however, that - depending on the samples $\mathcal{Z}_n$ - the problem (5.12) might not have a unique solution, i.e. G is not always invertible. In this case, we simply set $f_{\mathcal{Z}_n,V_k}(\mathbf{t}) := 0$ for all $\mathbf{t} \in T$.

Our goal is to provide a more refined result than the general estimates from chapter 4. To this end, we consider a truncated version of the regression problem. This can be seen as a substitute for the constant M-boundedness condition (4.35), which we used

for the constrained problem. Additionally, it is crucial to consider the unregularized and noiseless case to obtain a better rate than n^{-1} with respect to the number of samples.

First, we introduce the truncation operator and investigate its properties. Subsequently, we investigate the truncated version of the unregularized, noiseless regression problem (D) and derive upper bounds on the overall error in expectation.

5.4.1 The truncation operator

Definition 5.12 [Truncation Operator]
Let $\omega > 0$. We define the ***truncation operator*** $\tau_\omega : L_{\infty,\rho_T}(T;E) \to L_{\infty,\rho_T}(T;E)$ *by* $\tau_\omega(f) = P_\omega \circ f$, *where* $P_\omega : E \to E$ *is given by*

$$P_\omega(\mathbf{x}) := \begin{cases} \mathbf{x} & \text{if } \|\mathbf{x}\|_E \leq \omega, \\ \frac{\mathbf{x}}{\|\mathbf{x}\|_E} \cdot \omega & \text{else.} \end{cases} \tag{5.13}$$

Note that the image of the restriction of τ_ω to the $C(T;E)$ is a subset of $C(T;E)$.

The connection to M-boundedness

Let r be the bound from (4.5). Then, $r \geq \|g\|_{L_{\infty,\rho_T}(T;E)}$. Employing the truncation operator on functions from the search set V_k and considering $\tau_r(f)$ instead of $f \in V_k$ can be seen as a substitute for the constant M-boundedness condition $M_\psi \simeq 1$, see also (4.35), which was crucial to derive the optimal error bounds in section 4.5. To this end, note that

$$\|\tau_r(f)(\mathbf{t}) - \mathbf{x}\|_E^2 = \|P_r(f(\mathbf{t})) - \mathbf{x}\|_E^2 \leq (\|P_r(f(\mathbf{t}))\|_E + \|\mathbf{x}\|_E)^2 \leq (r+r)^2 = 4r^2 \simeq 1 \tag{5.14}$$

holds for ρ-almost every $(\mathbf{t}, \mathbf{x})$ and every $f \in V_k$. Therefore, employing the truncation operator τ_r on the search set V_k can be seen as a way to enforce $M_\psi \simeq 1$. However, it is noteworthy that $\tau_r(V_k)$ is not necessarily a subset of V_k.

We could also consider a truncated version of the more general constrained regression problems we tackled in chapter 4. However, the proofs for the estimates on the sampling error, which we provided in section 4.4, do not carry over to these truncated versions and we would have to come up with a different approach to establish theoretical convergence results there.

Contractiveness of the truncation operator

An important property of the truncation operator τ_ω is its contractiveness with respect to the $L_{p,\rho_T}(T;E)$ norm.

Lemma 5.13
Let $\omega > 0, p \in [1, \infty]$. The operator τ_ω is a contraction on $L_{p,\rho_T}(T; E)$, i.e. we have

$$\|\tau_\omega(f_1) - \tau_\omega(f_2)\|_{L_{p,\rho_T}(T;E)} \leq \|f_1 - f_2\|_{L_{p,\rho_T}(T;E)}$$

for $f_1, f_2 \in L_{\infty,\rho_T}(T; E)$.

Proof. Note that P_ω defined in (5.13) is the unique projection onto the convex ball of radius ω centered around $\mathbf{0}$ in E. Indeed, if there existed $\mathbf{x}, \mathbf{y} \in E$ with $\|\mathbf{x}\|_E > \omega$ and $\|\mathbf{y}\|_E \leq \omega$ such that $\|\mathbf{x} - \mathbf{y}\|_E < \|\mathbf{x} - P_\omega(\mathbf{x})\|_E$, then

$$\|\mathbf{x}\|_E \leq \|\mathbf{x} - \mathbf{y}\|_E + \|\mathbf{y}\|_E < \|\mathbf{x} - P_\omega(\mathbf{x})\|_E + \omega = \left(1 - \frac{\omega}{\|\mathbf{x}\|_E}\right) \|\mathbf{x}\|_E + w = \|\mathbf{x}\|_E,$$

which would be a contradiction. Therefore, P_ω is a contraction on E, see e.g. theorem 6.9 of [33] for details and a thorough proof of this fact. Since

$$\begin{aligned} \|\tau_\omega(f) - \tau_\omega(g)\|^p_{L_{p,\rho_T}(T;E)} &= \int_T \|P_\omega(f(\mathbf{t})) - P_\omega(g(\mathbf{t}))\|^p_E \, \mathrm{d}\rho_T(\mathbf{t}) \\ &\leq \int_T \|f(\mathbf{t}) - g(\mathbf{t})\|^p_E \, \mathrm{d}\rho_T(\mathbf{t}) = \|f - g\|^p_{L_{p,\rho_T}(T;E)} \end{aligned}$$

holds for $p \in [1, \infty)$, τ_ω is also a contraction on $L_{p,\rho_T}(T; E)$. For the remaining case $p = \infty$, we have

$$\begin{aligned} \|\tau_\omega(f) - \tau_\omega(g)\|_{L_{\infty,\rho_T}(T;E)} &= \operatorname*{ess\,sup}_{\mathbf{t} \in T} \|P_\omega(f(\mathbf{t})) - P_\omega(g(\mathbf{t}))\|_E \\ &\leq \operatorname*{ess\,sup}_{\mathbf{t} \in T} \|f(\mathbf{t}) - g(\mathbf{t})\|_E = \|f - g\|_{L_{\infty,\rho_T}(T;E)}, \end{aligned}$$

which completes the proof. □

5.4.2 An upper bound on the overall error

We now present an upper bound on the overall error for unregularized, noiseless function regression with finite-dimensional search spaces. To this end, we remind the reader of the $L_{2,\rho_T}(T; E)$-orthogonal projector $P_{V_k} : L_{2,\rho_T}(T; E) \to V_k$, see definition 4.16. Besides P_{V_k}, we also need the projector $P^n_{V_k}$ defined below.

Definition 5.14 [Orthogonal projection onto V_k with respect to $\|\cdot\|_{\mathcal{Z}_n}$]
We define the seminorm $\|\cdot\|_{\mathcal{Z}_n}$ on the space of all functions from T to E which are point evaluable in $\mathcal{Z}_n$ via the data-dependent scalar product

$$\langle f_1, f_2 \rangle_{\mathcal{Z}_n} := \frac{1}{n} \left\langle \begin{pmatrix} f_1(\mathbf{t}_1) \\ \vdots \\ f_1(\mathbf{t}_n) \end{pmatrix}, \begin{pmatrix} f_2(\mathbf{t}_1) \\ \vdots \\ f_2(\mathbf{t}_n) \end{pmatrix} \right\rangle_{E^n} = \frac{1}{n} \sum_{i=1}^n \langle f_1(\mathbf{t}_i), f_2(\mathbf{t}_i) \rangle_E.$$

If $\mathcal{Z}_n$ is such that the orthogonal projection onto V_k with respect to the $\|\cdot\|_{\mathcal{Z}_n}$ seminorm is well-defined, we denote this projection by $P^n_{V_k}$, i.e.

$$P^n_{V_k}(f) := \arg\min_{h\in V_k} \|h - f\|_{\mathcal{Z}_n} = \arg\min_{h\in V_k} \frac{1}{n}\sum_{i=1}^{n} \|h(\mathbf{t}_i) - f(\mathbf{t}_i)\|_E^2.$$

Note that $P^n_{V_k}$ is well-defined if G is invertible. To see this, we just need to replace g by f on the right hand side of the least-squares regression problem (5.12) to obtain the coefficients of $P^n_{V_k}(f)$. Indeed, considering the system

$$G\vec{\beta} = B(\overrightarrow{\boldsymbol{f}(\mathbf{t})}) \tag{5.15}$$

for a point-evaluable function $f : T \to E$ with $\overrightarrow{\boldsymbol{f}(\mathbf{t})} := (f(\mathbf{t}_1), \ldots, f(\mathbf{t}_n))^T$, we obtain $P^n_{V_k}(f) = \sum_{i=1}^n \beta_i \nu_i$. Note, furthermore, that $f_{\mathcal{Z}_n,V_k} = P^n_{V_k}(g)$. We are now in the position to formulate our theorem on the convergence of unregularized, noiseless function regression.

Theorem 5.15
Let $n \geq N_k$ and let $f_{\mathcal{Z}_n,V_k}$ be the solution to (D) *- or $f_{\mathcal{Z}_n,V_k} = 0$ if there is no unique solution to* (D)*. Let $K(N_k)$ and n be coupled such that they fulfill* (5.11) *for all $k \in \mathbb{N}$, i.e.*

$$K(N_k) \leq \frac{\lambda_{\min}(M)|\log(c_{\frac{1}{2}})|}{(1+\sigma)} \cdot \frac{n}{\log(n)}.$$

Then,

$$\begin{aligned}\mathbb{E}_{\rho_T^n}\left[\mathcal{E}\left(\tau_r\left(f_{\mathcal{Z}_n,V_k}\right)\right) - \mathcal{E}\left(g\right)\right] \leq \left(1 + \frac{4\lambda_{\max}(M)|\log(c_{\frac{1}{2}})|}{(1+\sigma)\lambda_{\min}(M)\log(n)}\right) \inf_{f\in V_k} \|f - g\|^2_{L_{2,\rho_T}(T;E)} \\ + 8r^2 n^{-\sigma},\end{aligned} \tag{5.16}$$

where the expectation has to be understood with respect to the product measure $\rho_T^n := \rho_T \times \ldots \times \rho_T$.

Proof. The general idea for the proof stems from [21]. However, we need to generalize it to the vector-valued setting and account for non-orthonormal bases. First, note that

$$\mathcal{E}\left(\tau_r\left(f_{\mathcal{Z}_n,V_k}\right)\right) - \mathcal{E}\left(g\right) = \|\tau_r\left(f_{\mathcal{Z}_n,V_k}\right) - g\|^2_{L_{2,\rho_T}(T;E)}$$

since $g = f_\rho$, see also lemma 4.4. Let $T^n = T \times \ldots \times T$ be the domain from which the n sample points $\mathbf{t}_i, i = 1, \ldots, n$ stem from. We distinguish two cases for the samples $\mathcal{Z}_n = (\mathbf{t}_i, g(\mathbf{t}_i))_{i=1}^n$. To this end, we split T^n into two parts: Let

$$T^n_+ := \left\{(\mathbf{t}_1, \ldots, \mathbf{t}_n) \in T^n \mid \|G\|_2 \leq \frac{3}{2}\lambda_{\max}(M) \text{ and } \|G^{-1}\|_2 \leq \frac{2}{\lambda_{\min}(M)}\right\}$$

and let $T_-^n := T^n \setminus T_+^n$. Note that T_-^n and T_+^n are ρ_T^n measurable since G and G^{-1} depend continuously on $\mathbf{t}_1, \dots, \mathbf{t}_n$. Due to lemma 5.10, we know that $\mathbb{P}_{\rho_T^n}(T_-^n) \leq 2n^{-\sigma}$. Now, let $E := \mathbb{E}_{\rho_T^n}\left[\|\tau_r\left(f_{\mathcal{Z}_n,V_k}\right) - g\|^2_{L_{2,\rho_T}(T;E)}\right]$. Because of (5.14), we have

$$\begin{aligned}
E &= \int_{T^n} \|\tau_r\left(f_{\mathcal{Z}_n,V_k}\right) - g\|^2_{L_{2,\rho_T}(T;E)} \,\mathrm{d}\rho_T^n \\
&= \int_{T_+^n} \|\tau_r\left(f_{\mathcal{Z}_n,V_k}\right) - g\|^2_{L_{2,\rho_T}(T;E)} \,\mathrm{d}\rho_T^n + \int_{T_-^n} \|\tau_r\left(f_{\mathcal{Z}_n,V_k}\right) - g\|^2_{L_{2,\rho_T}(T;E)} \,\mathrm{d}\rho_T^n \\
&\leq \int_{T_+^n} \|\tau_r\left(f_{\mathcal{Z}_n,V_k}\right) - g\|^2_{L_{2,\rho_T}(T;E)} \,\mathrm{d}\rho_T^n + \int_{T_-^n} 4r^2 \,\mathrm{d}\rho_T^n \\
&\leq \int_{T_+^n} \|\tau_r\left(f_{\mathcal{Z}_n,V_k}\right) - g\|^2_{L_{2,\rho_T}(T;E)} \,\mathrm{d}\rho_T^n + 8r^2 n^{-\sigma}. \qquad (5.17)
\end{aligned}$$

To estimate the first summand, note that $\tau_r(g) = g$ since $\|g\|_{L_{\infty,\rho_T}(T;E)} \leq r$ holds by definition of r. Therefore, we have

$$\|\tau_r\left(f_{\mathcal{Z}_n,V_k}\right) - g\|^2_{L_{2,\rho_T}(T;E)} = \|\tau_r\left(f_{\mathcal{Z}_n,V_k}\right) - \tau_r(g)\|^2_{L_{2,\rho_T}(T;E)} \leq \|f_{\mathcal{Z}_n,V_k} - g\|^2_{L_{2,\rho_T}(T;E)} \qquad (5.18)$$

because τ_r is a contraction on $L_{2,\rho_T}(T;E)$, see lemma 5.13. Furthermore, G is invertible on T_+^n and $P_{V_k}^n$ is well-defined. Therefore, $f_{\mathcal{Z}_n,V_k} = P_{V_k}^n(g)$ solves (D). Note that $P_{V_k}^n \circ P_{V_k} = P_{V_k}$ since the image of P_{V_k} is already an element of V_k. Thus, we have

$$\begin{aligned}
\|f_{\mathcal{Z}_n,V_k} - g\|^2_{L_{2,\rho_T}(T;E)} &= \|P_{V_k}^n(g) - P_{V_k}^n \circ P_{V_k}(g) + P_{V_k}(g) - g\|^2_{L_{2,\rho_T}(T;E)} \\
&= \|P_{V_k}^n\left(g - P_{V_k}(g)\right)\|^2_{L_{2,\rho_T}(T;E)} + \|g - P_{V_k}(g)\|^2_{L_{2,\rho_T}(T;E)} \qquad (5.19)
\end{aligned}$$

because $P_{V_k}^n$ is a linear operator and $g - P_{V_k}(g)$ is L_{2,ρ_T}-orthogonal on V_k. Using $f = g - P_{V_k}(g)$ in (5.15), we obtain $P_{V_k}^n\left(g - P_{V_k}(g)\right) = \sum_{j=1}^{N_k} \beta_j \nu_j$ for $\vec{\beta} = G^{-1}\vec{\xi}$ with

$$\xi_j := B(\overrightarrow{\boldsymbol{g - P_{V_k}(g)}(\mathbf{t})})_j = \frac{1}{n}\sum_{i=1}^{n} \langle \nu_j(\mathbf{t}_i), (g - P_{V_k}(g))(\mathbf{t}_i)\rangle_E$$

for each $j = 1, \dots, N_k$. Therefore, we get

$$\begin{aligned}
\|P_{V_k}^n\left(g - P_{V_k}(g)\right)\|^2_{L_{2,\rho_T}(T;E)} &= \vec{\beta}^T M \vec{\beta} = \vec{\xi}^T G^{-1} M G^{-1} \vec{\xi} \leq \|M\|_2 \, \|G^{-1}\|_2^2 \, \|\vec{\xi}\|_2^2 \\
&\leq \frac{4\lambda_{\max}(M)}{\lambda_{\min}(M)^2} \|\vec{\xi}\|_2^2 \qquad (5.20)
\end{aligned}$$

on T_+^n. To summarize what we have so far, we combine (5.17), (5.18), (5.19) and (5.20) to obtain

$$E \leq \int_{T_+^n} \left(\frac{4\lambda_{\max}(M)}{\lambda_{\min}(M)^2} \|\vec{\xi}\|_2^2 + \|g - P_{V_k}(g)\|^2_{L_{2,\rho_T}(T;E)} \right) \mathrm{d}\rho_T^n \; + \; 8r^2 n^{-\sigma},$$

which leads to

$$E \leq \frac{4\lambda_{\max}(M)}{\lambda_{\min}(M)^2} \cdot \mathbb{E}_{\rho_T^n}\left[\|\vec{\xi}\|_2^2\right] + \|g - P_{V_k}(g)\|^2_{L_{2,\rho_T}(T;E)} + 8r^2 n^{-\sigma} \tag{5.21}$$

since $\rho_T^n(T_+^n) \leq 1$. Because of the independence of $\mathbf{t}_1, \ldots, \mathbf{t}_n$, we observe that

$$\begin{aligned}
\mathbb{E}_{\rho_T^n}\left[\|\vec{\xi}\|_2^2\right] &= \int_{T^n} \sum_{j=1}^{N_k} \left(\frac{1}{n} \sum_{i=1}^{n} \langle \nu_j(\mathbf{t}_i), (g - P_{V_k}(g))(\mathbf{t}_i) \rangle_E \right)^2 \mathrm{d}\rho_T^n(\mathbf{t}_1, \ldots, \mathbf{t}_n) \\
&= \frac{1}{n^2} \sum_{j=1}^{N_k} \sum_{i,l=1}^{n} \int_{T\times T} \langle \nu_j(\mathbf{t}_i), (g - P_{V_k}(g))(\mathbf{t}_i) \rangle_E \\
&\qquad\qquad \cdot \langle \nu_j(\mathbf{t}_l), (g - P_{V_k}(g))(\mathbf{t}_l) \rangle_E \quad \mathrm{d}(\rho_T \times \rho_T)(\mathbf{t}_i, \mathbf{t}_j) \\
&= \frac{1}{n^2} \sum_{j=1}^{N_k} (n^2 - n) \left(\underbrace{\int_T \langle \nu_j(\mathbf{t}), (g - P_{V_k}(g))(\mathbf{t}) \rangle_E \, \mathrm{d}\rho_T(\mathbf{t})}_{=0} \right)^2 \\
&+ \frac{1}{n^2} \sum_{j=1}^{N_k} n \int_T \langle \nu_j(\mathbf{t}), (g - P_{V_k}(g))(\mathbf{t}) \rangle_E^2 \, \mathrm{d}\rho_T(\mathbf{t}) \\
&= \frac{1}{n} \sum_{j=1}^{N_k} \int_T \langle \nu_j(\mathbf{t}), (g - P_{V_k}(g))(\mathbf{t}) \rangle_E^2 \, \mathrm{d}\rho_T(\mathbf{t}),
\end{aligned}$$

where the last step follows from the L_{2,ρ_T}-orthogonality of $g - P_{V_k}(g)$ on V_k. By the definition of $K(N_k)$, we obtain

$$\begin{aligned}
\mathbb{E}_{\rho_T^n}\left[\|\vec{\xi}\|_2^2\right] &\leq \frac{1}{n} \int_T \sum_{j=1}^{N_k} \|\nu_j(\mathbf{t})\|_E^2 \|g - P_{V_k}(g)(\mathbf{t})\|_E^2 \, \mathrm{d}\rho_T(\mathbf{t}) \\
&\overset{(5.10)}{\leq} \frac{1}{n} K(N_k) \|g - P_{V_k}(g)\|^2_{L_{2,\rho_T}(T;E)} \\
&\overset{(5.11)}{\leq} \frac{\lambda_{\min}(M) |\log(c_{\frac{1}{2}})|}{(1+\sigma)\log(n)} \|g - P_{V_k}(g)\|^2_{L_{2,\rho_T}(T;E)}.
\end{aligned}$$

Applying this to (5.21), we finally obtain

$$E \leq \frac{4\lambda_{\max}(M) |\log(c_{\frac{1}{2}})|}{(1+\sigma)\lambda_{\min}(M)\log(n)} \|g - P_{V_k}(g)\|^2_{L_{2,\rho_T}(T;E)} + \|g - P_{V_k}(g)\|^2_{L_{2,\rho_T}(T;E)} + 8r^2 n^{-\sigma},$$

which completes the proof since $\inf_{f \in V_k} \|f - g\|_{L_{2,\rho_T}(T;E)} = \|g - P_{V_k}(g)\|_{L_{2,\rho_T}(T;E)}$. □

Under the condition that the coupling between the discretization scale k and the number of samples n fulfills (5.11), theorem 5.15 provides us with an upper bound on

the expectation of the overall error for a truncated, unregularized, noiseless regression method which operates on finite-dimensional search spaces. We now consider the result of theorem 5.15 in more detail and give a few remarks:

- The decomposition of the error (5.16) into two summands reminds us of the decomposition into the bias and the sampling error from the previous chapter. However, the first summand of (5.16) essentially[4] only depends on the discretization scale k, whereas the second summand only depends on n. Therefore, no coupling is present in each of the summands. This is different to theorem 4.24, where the sampling error is influenced by both, the size of V_k and the number of samples n.

- Inspecting the first summand of (5.16) in detail, we see that the prefactor

$$1 + \frac{4\lambda_{\max}(M)|\log(c_{\frac{1}{2}})|}{(1+\sigma)\lambda_{\min}(M)\log(n)}$$

  in front of the best approximation error converges to 1 with $n \to \infty$ if the condition number $\kappa(M) = \frac{\lambda_{\max}(M)}{\lambda_{\min}(M)}$ is bounded, i.e. if we are dealing with a Riesz basis. The second summand of (5.16) shows that - depending on how large $\sigma > 0$ can be chosen such that (5.11) is still fulfilled - the convergence rate for $n \to \infty$ can indeed be faster than n^{-1}, which was the best rate we could achieve in chapter 4. We will see this in more detail in section 5.5.

- It is noteworthy that a similar result can be derived also for the noisy case of unregularized function regression, i.e. $\mathbf{x}_i = g(\mathbf{t}_i) + \varepsilon_i$, where the noise distribution has finite variance. Here, basically a third summand which scales like $\frac{\mathrm{Var}\cdot N_k}{n}$ has to be added to (5.16), where Var denotes the variance of the noise distribution, see [21] for an analysis of the scalar-valued case with orthonormal bases. In this thesis, we neglect the thorough analysis of this setting as it does not lead to significant additional insights compared to the results of chapter 4, where we also observed the rate $\frac{N_k}{n}$ for the sampling error up to logarithms if $M_\psi \simeq 1$, see e.g. (4.52) or (4.57).

- There also exist techniques to derive probabilistic error bounds in the setting of noiseless function regression, which still exhibit essentially the same structure as (5.16). Several results for different noise models for scalar-valued function regression can be found in [60].

Summary

Based on the stability results of lemma 5.10 and theorem 5.11, we derived a bound on the expected error for unregularized, noiseless function regression in this section. Our

[4] Note that the $\frac{1}{\log(n)}$ can be neglected in this summand as it gets smaller for $n \to \infty$ and the rate is therefore mainly determined by the best approximation error in V_k.

main findings are the following:

- The expected $L_{2,\rho_T}(T;E)$ error when computing the truncated solution $\tau_r(f_{\mathcal{Z}_n,V_k})$ of the unregularized, noiseless function regression problem (D), instead of taking the true function g, is

$$\left(1+\frac{4\lambda_{\max}(M)|\log(c_{\frac{1}{2}})|}{(1+\sigma)\lambda_{\min}(M)\log(n)}\right)\inf_{f\in V_k}\|f-g\|^2_{L_{2,\rho_T}(T;E)}+8r^2n^{-\sigma}$$

  if the discretization scale k and the number of samples n are coupled such that $n \geq N_k$ and

$$K(N_k) \leq \frac{\lambda_{\min}(M)|\log(c_{\frac{1}{2}})|}{(1+\sigma)}\cdot\frac{n}{\log(n)}$$

  are fulfilled. The first summand in the error term resembles the choice of the finite-dimensional search space V_k and its basis $\nu_1,\ldots,\nu_{N_k}$. It is mainly governed by the condition number $\kappa(M)$ of the mass matrix and the $L_{2,\rho_T}(T;E)$ best approximation error in V_k. The second summand represents the error due to the finite sample size.

- The first summand of the error (5.16) behaves comparably to the discretization error in theorem 4.18. The second summand, however, now decays faster than n^{-1} if σ can be chosen larger than 1. Therefore, if we are dealing with noiseless function regression, we can improve on the sampling error bound from theorem 4.24.

5.5 Examples for noiseless function regression

Analogously to section 4.5, we now consider the overall error in noiseless function regression for specific choices of V_k. As in the previous chapter, we have a look at linear spline spaces on full grids and sparse grids, as well as Fourier polynomials on full grids and hyperbolic crosses.

Note that the Fourier bases of $\mathcal{T}_k^{\text{full},d}$ and $\mathcal{T}_k^{\text{hyp},d}$ are L_2-orthonormal. Furthermore, the prewavelet bases of $\mathcal{V}_k^{\text{full},d}$ and $\mathcal{V}_k^{\text{sparse},d}$ are L_2 Riesz bases in the sense that

$$\vec{\beta}^T M\vec{\beta} \simeq \|\vec{\beta}\|_2^2$$

holds independently of $k \in \mathbb{N}$ for the corresponding mass matrices M and arbitrary coefficient vectors $\vec{\beta} \in \mathbb{R}^{N_k}$. This is due to the norm equivalences (3.37) and (3.38) for $s = 0$. Therefore, the condition number $\kappa(M) = \frac{\lambda_{\max}(M)}{\lambda_{\min}(M)}$ is bounded from above independently of $k \in \mathbb{N}$ for every search space V_k which we study here. As we mentioned in section 5.3, this means that (5.11) suffices to ensure the stability of the regression problem with high probability, see also theorem 5.11.

The condition $n \geq N_k$ is a necessary prerequisite in both, theorem 5.11, and theorem 5.15. Therefore, it is a necessary condition to obtain stability and convergence results

for all of our examples. Our approach will be the same for each example:

- First, we study the behavior of $K(N_k)$ from (5.10).
- We use the scaling of $K(N_k)$ to determine a coupling between k and n which fulfills (5.11) and, thus, gives a sufficient condition to ensure stability of the regression method with high probability.
- We apply theorem 5.15 to obtain a bound on the expected convergence rate for unregularized, noiseless function regression.
- Finally, we approximately balance the two summands of the error term (5.16) to derive an optimal coupling between the basis size N_k and the number of samples n. We also determine the appropriate oversampling constant $\sigma > 0$ and present the convergence rate with respect to n in the balanced case.

After the analysis of our examples, we give a short overview on our findings and conclude this section by relating our results to the work of other researchers.

5.5.1 Regression with piecewise linear basis functions on full grids

We again have a look at the setting from subsection 4.5.1 of multivariate regression on full grids with piecewise linear basis functions. However, we now focus on the case of noiseless function regression (D), i.e. $T = (0,1)^m$, $\rho_T = \lambda_T$ and $f_\rho = g$ with $g \in H^s(T; \mathbb{R}^d)$ for a fixed $0 < s \leq 2$. The search spaces are the prewavelet spaces $V_k = \mathcal{V}_k^{\text{full},d}$ on a full grid of level $k \in \mathbb{N}$.

Stability

To apply theorem 5.11, we need to establish an upper bound on $K(N_k)$. Since the basis functions of V_k are built componentwise, see (3.35), and because of the definition of $K(N_k)$ in (5.10), we have

$$K(N_k) = d \cdot \sup_{\mathbf{t} \in T} \sum_{|\mathbf{l}|_{\ell_\infty} \leq k} \sum_{\mathbf{i} \in \mathbf{I}_\mathbf{l}} \gamma_{\mathbf{l},\mathbf{i}}(\mathbf{t})^2.$$

In order to obtain an upper bound on this quantity, we first need an auxiliary result.

Lemma 5.16
For each $\mathbf{l} \in \mathbb{N}^m$, we have

$$\max_{\mathbf{t} \in [0,1]^m} \sum_{\mathbf{i} \in \mathbf{I}_\mathbf{l}} \gamma_{\mathbf{l},\mathbf{i}}(\mathbf{t})^2 \leq 2^{|\mathbf{l}|_{\ell_1}} \cdot 2^{|\{j \in \{1,\ldots,m\} \,|\, l_j = 0\}|} \cdot \left(\frac{36}{25}\right)^{|\{j \in \{1,\ldots,m\} \,|\, l_j > 0\}|}. \tag{5.22}$$

Proof. We start with the univariate case $m = 1$. Let us shortly recall the definition of the univariate prewavelet basis functions $\gamma_{l,i}$ on $[0,1]$ from subsection 3.5.1: $\gamma_{0,0} := 1$, $\gamma_{0,1}(t) := t$, $\gamma_{1,1} := 2 \cdot \phi_{1,1} - 1$. For $l \geq 2$, we have

$$\gamma_{l,i} := 2^{\frac{l}{2}} \cdot \left(\frac{1}{10}\phi_{l,i-2} - \frac{6}{10}\phi_{l,i-1} + \phi_{l,i} - \frac{6}{10}\phi_{l,i+1} + \frac{1}{10}\phi_{l,i+2} \right)$$

for $i \in I_l, i \neq 1, 2^l - 1$ and

$$\gamma_{l,1} := 2^{\frac{l}{2}} \cdot \left(-\frac{6}{5}\phi_{l,0} + \frac{11}{10}\phi_{l,1} - \frac{3}{5}\phi_{l,2} + \frac{1}{10}\phi_{l,3} \right), \; \gamma_{l,2^l-1}(t) := \gamma_{l,1}(1-t),$$

where the hat functions $\phi_{l,i}$ take values between 0 and 1.

We define $S_l(t) := \sum_{i \in I_l} \gamma_{l,i}(t)^2$. For the special case $l = 0$, we obtain

$$S_0(t) = \gamma_{0,0}^2(t) + \gamma_{0,1}^2(t) = 1 + t^2 \leq 2$$

and for $l = 1$ we have

$$S_1(t) = \gamma_{1,1}^2(t) = (2\phi_{1,1}(t) - 1)^2 \leq 1$$

with $t \in [0,1]$. For the case $l \geq 2$, note that S_l is the sum of the piecewise quadratic polynomials $\gamma_{l,i}^2(\cdot)$ with $i \in I_l$. Between consecutive grid nodes, the coefficient of the quadratic term of $\gamma_{l,i}^2(\cdot)$ is always positive for every $i \in I_l$. Therefore, the quadratic term of the piecewise quadratic polynomial $S_l(\cdot)$ also has a positive coefficient everywhere. This shows that the maximum of S_l over $[0,1]$ can only reside on one of the grid points $2^{-l}i$ with $i = 0, \ldots, 2^l$. This can also be seen from the example in figure 5.1, where S_4 is plotted.

We claim that the maximum of S_l is attained at the boundary point $t = 1$. For $l = 0$ and $l = 1$, this is directly clear. Let us consider the values at the grid nodes for S_2. These follow directly from the definition of the univariate prewavelet functions. We use a mask-type notation which contains a prefactor 2^l and the nodal values at the grid points. The calculation

$$\begin{aligned}
S_2(t) &= \gamma_{2,1}^2(t) + \gamma_{2,3}^2(t) \\
&= 4 \; [\; \tfrac{36}{25} \quad \tfrac{121}{100} \quad \tfrac{9}{25} \quad \tfrac{1}{100} \quad 0 \quad] \\
&+ 4 \; [\; 0 \quad \tfrac{1}{100} \quad \tfrac{9}{25} \quad \tfrac{121}{100} \quad \tfrac{36}{25} \quad] \\
&= 4 \; [\; \tfrac{36}{25} \quad \tfrac{61}{50} \quad \tfrac{18}{25} \quad \tfrac{61}{50} \quad \tfrac{36}{25} \quad]
\end{aligned}$$

shows that the largest value $4 \cdot \frac{36}{25}$ is attained at the boundary grid points. Analogously,

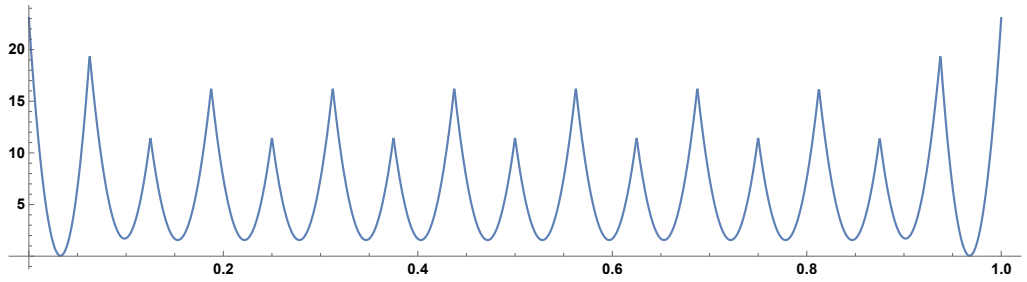

Fig. 5.1: The squared sum S_4 of the univariate prewavelet basis functions on level $k = 4$.

we have

$$
\begin{aligned}
S_3(t) &= \gamma_{3,1}^2(t) + \gamma_{3,3}^2(t) + \gamma_{3,5}^2(t) + \gamma_{3,7}^2(t) \\
&= 8 \left[\tfrac{36}{25} \;\; \tfrac{121}{100} \;\; \tfrac{9}{25} \;\; \tfrac{1}{100} \;\; 0 \;\; 0 \;\; 0 \;\; 0 \;\; 0 \right] \\
&+ 8 \left[0 \;\; \tfrac{1}{100} \;\; \tfrac{9}{25} \;\; 1 \;\; \tfrac{9}{25} \;\; \tfrac{1}{100} \;\; 0 \;\; 0 \;\; 0 \right] \\
&+ 8 \left[0 \;\; 0 \;\; 0 \;\; \tfrac{1}{100} \;\; \tfrac{9}{25} \;\; 1 \;\; \tfrac{9}{25} \;\; \tfrac{1}{100} \;\; 0 \right] \\
&+ 8 \left[0 \;\; 0 \;\; 0 \;\; 0 \;\; 0 \;\; \tfrac{1}{100} \;\; \tfrac{9}{25} \;\; \tfrac{121}{100} \;\; \tfrac{36}{25} \right] \\
&= 8 \left[\tfrac{36}{25} \;\; \tfrac{61}{50} \;\; \tfrac{18}{25} \;\; \tfrac{51}{50} \;\; \tfrac{18}{25} \;\; \tfrac{51}{50} \;\; \tfrac{18}{25} \;\; \tfrac{61}{50} \;\; \tfrac{36}{25} \right]
\end{aligned}
$$

for $l = 3$. For higher levels l, these calculations work completely analogously. Due to the local support and overlap of the basis functions, the values of S_l cannot exceed $2^l \cdot \frac{36}{25}$ also for higher levels l. Therefore, for each $l \in \mathbb{N}$, the maximum of S_l is attained for $t = 1$. If $l = 0$, this maximum is 2 and if $l \geq 2$, this maximum is $2^l \cdot \frac{36}{25}$. In the special case $l = 1$, we just take the crude estimate $S_1(1) = 1 < 2 \cdot \frac{36}{25}$. Therefore, the assertion (5.22) is proven for $m = 1$.

The higher-dimensional case $m > 1$ follows due to the tensor product construction of the basis functions. To this end, let $\mathbf{t} \in [0,1]^m$ and $\mathbf{l} \in \mathbb{N}^m$ be arbitrary. Note that

$$
\sum_{\mathbf{i} \in \mathbf{I_l}} \gamma_{\mathbf{l},\mathbf{i}}(\mathbf{t})^2 = \sum_{(i_1,\ldots,i_m) \in \mathbf{I_l}} \prod_{j=1}^{m} \gamma_{l_j,i_j}(t_j)^2 = \prod_{j=1}^{m} \sum_{i_j \in I_{l_j}} \gamma_{l_j,i_j}(t_j)^2
$$

holds due to the structure of $\mathbf{I_l}$. Therefore, the maximization of the whole term can be split into the maximization of S_{l_j} for each direction $j \in \{1,\ldots,m\}$. Since the maximum is bounded by 2 for directions j with $l_j = 0$ and it is bounded by $2^{l_j} \cdot \frac{36}{25}$ for directions j with $l_j \geq 1$, the claimed inequality (5.22) follows. □

With the help of lemma 5.16 and using $\frac{36}{25} < 2$, we can now estimate $K(N_k)$ by

$$\begin{aligned} K(N_k) &\leq d \cdot \sum_{|\mathbf{l}|_{\ell_\infty} \leq k} 2^{|\mathbf{l}|_{\ell_1}} \cdot 2^{|\{j \in \{1,\dots,m\} \,|\, l_j = 0\}|} \cdot \left(\frac{36}{25}\right)^{|\{j \in \{1,\dots,m\} \,|\, l_j > 0\}|} \leq 2^m d \sum_{|\mathbf{l}|_{\ell_\infty} \leq k} 2^{|\mathbf{l}|_{\ell_1}} \\ &= 2^m \cdot d \cdot \prod_{j=1}^{m} \sum_{0 \leq l_j \leq k} 2^{l_j} = 2^m \cdot d \cdot \left(\sum_{0 \leq l \leq k} 2^l \right)^m = 2^m \cdot d \left(\frac{1 - 2^{k+1}}{1 - 2} \right)^m \\ &= 2^m \cdot d \cdot (2^{k+1} - 1)^m \leq 2^m \cdot d \cdot 2^{(k+1)m} < 4^m \cdot N_k, \end{aligned}$$

where we used $N_k = d \cdot (2^k + 1)^m > d \cdot 2^{km}$.

With this estimate, the condition (5.11) on $K(N_k)$, which guarantees stability of noiseless function regression with probability larger than $1 - 2n^{-\sigma}$, see theorem 5.11, becomes

$$4^m N_k \leq \frac{\lambda_{\min}(M) |\log(c_{\frac{1}{2}})|}{(1+\sigma)} \cdot \frac{n}{\log(n)}. \tag{5.23}$$

Since we are dealing with a Riesz basis, $\lambda_{\min}(M)$ is bounded from below and the sufficient scaling between n and k to obtain stability with high probability is essentially

$$2^{km} \simeq N_k \lesssim \frac{n}{\log(n)}, \tag{5.24}$$

where $\simeq$ and $\lesssim$ imply constants depending on m, d and σ.

The overall error

If the condition (5.23) is fulfilled, we can apply theorem 5.15 to obtain a bound on the expected overall error $E := \mathbb{E}_{\lambda_T^n} \left[\mathcal{E}\left(\tau_r\left(f_{\mathcal{Z}_n, V_k}\right)\right) - \mathcal{E}\left(g\right) \right]$ of the noiseless regression problem with $V_k = \mathcal{V}_k^{\text{full},d}$ and $g \in H^s(T; \mathbb{R}^d)$ with $0 < s \leq 2$, namely

$$\begin{aligned} E \;&\leq \left(1 + \frac{4\lambda_{\max}(M) |\log(c_{\frac{1}{2}})|}{(1+\sigma)\lambda_{\min}(M)\log(n)}\right) \inf_{f \in V_k} \|f - g\|^2_{L_2(T;\mathbb{R}^d)} + 8r^2 n^{-\sigma} \\ &\lesssim \inf_{f \in V_k} \|f - g\|^2_{L_2(T;\mathbb{R}^d)} + n^{-\sigma} \overset{(3.41)}{\lesssim} d 2^{-2sk} \|g\|^2_{H^s(T;\mathbb{R}^d)} + n^{-\sigma} \lesssim 2^{-2sk} + n^{-\sigma}. \end{aligned} \tag{5.25}$$

Here, the constant in the final $\lesssim$ estimate depends on m, d, σ, r, s and $\|g\|_{H^s(T;\mathbb{R}^d)}$. The expectation has to be understood with respect to the n-fold product λ_T^n of the Lebesgue measure λ_T on $T = (0,1)^m$.

Balancing the error terms

In order to fulfill the stability condition with the least amount of data points, we first choose the largest k (in dependence on σ and n) such that (5.23) still holds. Thus, the

inequality in (5.24) becomes an equality and we have

$$2^{km} \simeq N_k \simeq \frac{n}{\log(n)}$$

in this case. Substituting this into (5.25), we obtain

$$E \lesssim 2^{-2sk} + n^{-\sigma} \simeq \left(2^{km}\right)^{-\frac{2s}{m}} + n^{-\sigma} \simeq \left(\frac{\log(n)}{n}\right)^{\frac{2s}{m}} + n^{-\sigma}.$$

Therefore, we see that the overall error is approximately balanced for $\sigma = \frac{2s}{m}$ and we obtain the overall rate

$$E = \mathcal{O}\left(n^{-\frac{2s}{m}} \log(n)^{\frac{2s}{m}}\right).$$

5.5.2 Regression with piecewise linear basis functions on sparse grids

Now we consider the setting from subsection 4.5.2 for multivariate, noiseless sparse grid regression with piecewise linear prewavelets. To this end, let $T = (0,1)^m$, $\rho_T = \lambda_T$ and $f_\rho = g \in H^s_{\text{mix}}(T; \mathbb{R}^d)$ for a fixed $0 < s \leq 2$. We employ $V_k = \mathcal{V}_k^{\text{sparse},d}$ with $k \in \mathbb{N}$ as search spaces.

Stability

We recall the definition $\zeta_m(\mathbf{l}) := |\mathbf{l}|_{\ell_1} - m + |\{j \in \{1,\dots,m\} \mid l_j = 0\}| + 1$ from subsection 3.5.1. Similar as in the full grid example, we now have to estimate

$$K(N_k) = d \cdot \sup_{\mathbf{t} \in T} \sum_{\zeta_m(\mathbf{l}) \leq k} \sum_{\mathbf{i} \in \mathbf{I}_\mathbf{l}} \gamma_{\mathbf{l},\mathbf{i}}(\mathbf{t})^2.$$

Lemma 5.17
For $V_k = \mathcal{V}_k^{sparse,d}$, $K(N_k)$ can be bounded by

$$K(N_k) \leq 2 \cdot \left(\frac{72}{25}\right)^m N_k. \tag{5.26}$$

Proof. We denote the number of zero indices of a multiindex $\mathbf{l} \in \mathbb{N}^m$ by $Z(\mathbf{l}) := |\{j \in \{1,\dots,m\} \mid l_j = 0\}|$. By applying lemma 5.16, we get

$$K(N_k) \leq d \cdot \sum_{|\mathbf{l}|_{\ell_1} + Z(\mathbf{l}) \leq k+m-1} 2^{|\mathbf{l}|_{\ell_1} + Z(\mathbf{l})} \cdot \left(\frac{36}{25}\right)^{m - Z(\mathbf{l})},$$

where we used $\zeta_m(\mathbf{l}) = |\mathbf{l}|_{\ell_1} - m + Z(\mathbf{l}) + 1$. Substituting $i = |\mathbf{l}|_{\ell_1} + Z(\mathbf{l})$, the estimate

can be reformulated as

$$K(N_k) \leq d \sum_{i=0}^{k+m-1} 2^i \cdot \sum_{l=0}^{m} |\{\mathbf{l} \in \mathbb{N}^m \mid |\mathbf{l}|_{\ell_1} = i - l \text{ and } Z(\mathbf{l}) = l\}| \cdot \left(\frac{36}{25}\right)^{m-l}.$$

Note that $|\{\mathbf{l} \in \mathbb{N}^m \mid |\mathbf{l}|_{\ell_1} = i - l \text{ and } Z(\mathbf{l}) = l\}| = 0$ for all $l = 0, \ldots, m$ if $i < m$. Therefore, we can start the summation over i from m. If $i \geq m$, simple combinatorial arguments, see also [14] and the proof of lemma 3.23, lead to

$$\begin{aligned} |\{\mathbf{l} \in \mathbb{N}^m \mid |\mathbf{l}|_{\ell_1} = i - l \text{ and } Z(\mathbf{l}) = l\}| &= |\{\mathbf{l} \in (\mathbb{N} \setminus \{0\})^{m-l} \mid |\mathbf{l}|_{\ell_1} = i - l\}| \cdot \binom{m}{l} \\ &= \binom{i-l-1}{m-l-1}\binom{m}{l} \end{aligned}$$

for any $l = 0, \ldots, m-1$ and

$$|\{\mathbf{l} \in \mathbb{N}^m \mid |\mathbf{l}|_{\ell_1} = i - m \text{ and } Z(\mathbf{l}) = m\}| = \begin{cases} 1 & \text{if } i = m \\ 0 & \text{else} \end{cases} = \delta_{im}$$

in the case $l = m$. Therefore, we obtain

$$\begin{aligned} K(N_k) &\leq d \sum_{i=m}^{k+m-1} 2^i \cdot \left(\delta_{im} + \sum_{l=0}^{m-1} \binom{i-l-1}{m-l-1}\binom{m}{l}\left(\frac{36}{25}\right)^{m-l}\right) \\ &= d \cdot 2^m \cdot \sum_{i=0}^{k-1} 2^i \cdot \left(\delta_{i0} + \sum_{l=0}^{m-1} \binom{i+m-l-1}{m-l-1}\binom{m}{l}\left(\frac{36}{25}\right)^{m-l}\right) \\ &= d \cdot 2^m \cdot \left(1 + \sum_{l=0}^{m-1} \left(\frac{36}{25}\right)^{m-l} \binom{m}{l}\left(\sum_{i=0}^{k-1} 2^i \binom{i+m-l-1}{m-l-1}\right)\right) \\ &= d \cdot 2^m + 2^m \sum_{l=0}^{m-1} \left(\frac{36}{25}\right)^{m-l} \binom{m}{l} |G_k^{m-l}|, \end{aligned}$$

where $|G_k^{m-l}|$ denotes the size of a sparse grid of level k in dimension $m - l$ without boundary points, see lemma 3.6 of [14] for a proof of the last equality in the case $d = 1$. The vector-valued case $d > 1$ follows directly since we use a sparse grid of the same size in every component. To derive a bound with respect to N_k, we rewrite the above inequality to obtain

$$\begin{aligned} K(N_k) &\leq d \cdot 2^m + \sum_{l=0}^{m-1} \left(2 \cdot \frac{36}{25}\right)^{m-l} \cdot 2^l \binom{m}{l} |G_k^{m-l}| \\ &\leq d \cdot 2^m + \left(\frac{72}{25}\right)^m \cdot \sum_{l=0}^{m-1} 2^l \binom{m}{l} |G_k^{m-l}| = d \cdot 2^m + \left(\frac{72}{25}\right)^m N_k, \end{aligned}$$

where the last equality is proven in lemma 2.1.2 of [30]. The fact that $d \cdot 2^m \leq \left(\frac{72}{25}\right)^m N_k$ holds for each $k > 0$ completes the proof. □

Lemma 5.17 and theorem 5.11 show that unregularized, noiseless function regression is stable for the prewavelet sparse grid spaces V_k with probability larger than $1 - 2n^{-\sigma}$ if

$$2 \cdot \left(\frac{72}{25}\right)^m N_k \leq \frac{\lambda_{\min}(M) n |\log(c_{\frac{1}{2}})|}{(1+\sigma)\log(n)}. \tag{5.27}$$

Since the prewavelets in V_k form a Riesz basis with respect to $L_2(T;\mathbb{R}^d)$, the condition number $\kappa(M) = \frac{\lambda_{\max}(M)}{\lambda_{\min}(M)}$ of the mass matrix is bounded from above independently of $k \in \mathbb{N}$ and the scaling between k and n essentially becomes

$$2^k k^{m-1} \simeq N_k \lesssim \frac{n}{\log(n)}, \tag{5.28}$$

with implicit m, d and σ dependent constants.

The overall error

We assume that (5.27) is fulfilled and apply theorem 5.15 to obtain the expected convergence rate

$$E \lesssim \inf_{f \in V_k} \|f - g\|^2_{L_2(T;\mathbb{R}^d)} + n^{-\sigma} \overset{(3.42)}{\lesssim} d 2^{-2sk} k^{m-1} \|g\|^2_{H^s_{\text{mix}}(T;\mathbb{R}^d)} + n^{-\sigma} \lesssim 2^{-2sk} k^{m-1} + n^{-\sigma} \tag{5.29}$$

for the error $E := \mathbb{E}_{\lambda_T^n}\left[\mathcal{E}\left(\tau_r\left(f_{\mathcal{Z}_n, V_k}\right)\right) - \mathcal{E}\left(g\right)\right]$ of the noiseless regression problem (D) with $V_k = \mathcal{V}_k^{\text{sparse},d}$ and $g \in H^s_{\text{mix}}(T;\mathbb{R}^d)$ with $0 < s \leq 2$. Similarly to the full grid case, the constant in the final $\lesssim$ estimate depends on m, d, σ, r, s and $\|g\|_{H^s_{\text{mix}}(T;\mathbb{R}^d)}$.

Balancing the error terms

Let k be the largest natural number such that (5.27) is fulfilled. According to (5.28), we then have

$$2^k k^{m-1} \simeq N_k \simeq \frac{n}{\log(n)}.$$

For $n > 1$, we can take the logarithm on both sides and obtain $k + (m-1)\log(k) \simeq \log(n) - \log(\log(n))$, which leads to $k \lesssim \log(n)$. Therefore, by substituting our above coupling of k and n into (5.29), we get

$$\begin{aligned} E &\lesssim 2^{-2sk} k^{m-1} + n^{-\sigma} \simeq \left(2^k k^{m-1}\right)^{-2s} k^{(2s+1)(m-1)} + n^{-\sigma} \\ &\lesssim \left(\frac{\log(n)}{n}\right)^{2s} \log(n)^{(2s+1)(m-1)} + n^{-\sigma} \simeq n^{-2s} \log(n)^{(2s+1)m-1} + n^{-\sigma}, \end{aligned}$$

which is approximately balanced for $\sigma = 2s$. In this case, we obtain the overall rate

$$E = \mathcal{O}\left(n^{-2s}\log(n)^{(2s+1)m-1}\right).$$

5.5.3 Periodic regression with Fourier polynomials on full grids

To investigate noiseless function regression in the periodic setting, we assume that $T = (-\pi,\pi)^m, \rho_T = \frac{1}{(2\pi)^m}\lambda_T$ and $g \in \bar{H}^s(T;\mathbb{C}^d)$ for an $s > 0$. Here, we consider Fourier polynomials on full grids. To this end, we take $V_k = \mathcal{T}_k^{\text{full},d}$ for all $k \in \mathbb{N}$.

Stability

For the Fourier basis, we directly see that

$$K(N_k) = d \cdot \sup_{\mathbf{t}\in T} \sum_{\substack{\mathbf{l}\in\mathbb{Z}^m \\ |\mathbf{l}|_\infty \le 2^k}} |\mathrm{e}^{i\mathbf{l}^T\mathbf{t}}|_{\mathbb{C}}^2 = d \cdot \sum_{\substack{\mathbf{l}\in\mathbb{Z}^m \\ |\mathbf{l}|_\infty \le 2^k}} 1 = N_k. \tag{5.30}$$

Therefore, theorem 5.11 states that solving the noiseless problem (D) is stable with probability $1 - 2n^{-\sigma}$ if

$$N_k = K(N_k) \le \frac{\lambda_{\min}(M)|\log(c_{\frac{1}{2}})|}{(1+\sigma)} \cdot \frac{n}{\log(n)}. \tag{5.31}$$

Since the Fourier basis is $L_{2,\rho_T}(T;\mathbb{C}^d)$-orthonormal, the mass matrix M is the identity matrix for each $k \in \mathbb{N}$, i.e. $\lambda_{\max}(M) = \lambda_{\min}(M) = 1$ and the scaling in the above condition becomes

$$2^{mk} \simeq N_k \lesssim \frac{n}{\log(n)}. \tag{5.32}$$

The overall error

If (5.31) holds, the error $E := \mathbb{E}_{\rho_T^n}\left[\mathcal{E}\left(\tau_r\left(f_{\mathcal{Z}_n,V_k}\right)\right) - \mathcal{E}\left(g\right)\right]$ fulfills

$$E \lesssim \inf_{f\in V_k} \|f-g\|_{L_2(T;\mathbb{C}^d)}^2 + n^{-\sigma} \overset{(3.50)}{\lesssim} d2^{-2sk}\|g\|_{\bar{H}^s(T;\mathbb{C}^d)}^2 + n^{-\sigma} \lesssim 2^{-2sk} + n^{-\sigma} \tag{5.33}$$

with an implicit constant depending on m, d, σ, r, s and $\|g\|_{\bar{H}^s(T;\mathbb{C}^d)}$, according to theorem 5.15.

Balancing the error terms

Depending on n and σ, we choose the largest k such that (5.31) is still fulfilled. Then, we obtain the approximate equality

$$2^{mk} \simeq N_k \simeq \frac{n}{\log(n)},$$

which we substitute into (5.33). This leads to

$$E \lesssim 2^{-2sk} + n^{-\sigma} \simeq \left(2^{mk}\right)^{-\frac{2s}{m}} + n^{-\sigma} \simeq \left(\frac{\log(n)}{n}\right)^{\frac{2s}{m}} + n^{-\sigma}.$$

We see that $\sigma = \frac{2s}{m}$ is the right choice to balance these terms and get

$$E = \mathcal{O}\left(n^{-\frac{2s}{m}} \log(n)^{\frac{2s}{m}}\right).$$

5.5.4 Periodic regression with Fourier polynomials on hyperbolic crosses

Finally, we consider noiseless function regression in the periodic setting with Fourier polynomials on hyperbolic crosses. To this end, we again take $T = (-\pi, \pi)^m$ and $\rho_T = \frac{1}{(2\pi)^m}\lambda_T$. We assume that $s > 0$ is such that $g \in \bar{H}^s_{\mathrm{mix}}(T; \mathbb{C}^d)$ and take $V_k = \mathcal{T}_k^{\mathrm{hyp},d}$ for all $k \in \mathbb{N}$.

Stability

With the same argument as in (5.30), we obtain

$$K(N_k) = d \cdot \sup_{\mathbf{t}\in T} \sum_{\substack{\mathbf{l}\in\mathbb{Z}^m \\ \prod_{n=1}^m (\max(|l_n|,1)) \le 2^k}} |\mathrm{e}^{i\mathbf{l}^T\mathbf{t}}|^2_{\mathbb{C}} = d \cdot \sum_{\substack{\mathbf{l}\in\mathbb{Z}^m \\ \prod_{n=1}^m (\max(|l_n|,1)) \le 2^k}} 1 = N_k.$$

Thus, theorem 5.11 shows that solving (D) over V_k is stable with probability $1 - 2n^{-\sigma}$ if

$$N_k = K(N_k) \le \frac{\lambda_{\min}(M)|\log(c_{\frac{1}{2}})|}{(1+\sigma)} \cdot \frac{n}{\log(n)}. \tag{5.34}$$

Again, due to the $L_{2,\rho_T}(T; \mathbb{C}^d)$-orthonormality of the Fourier basis, we get $\lambda_{\max}(M) = \lambda_{\min}(M) = 1$ independently of $k \in \mathbb{N}$. Hence, the scaling which is necessary to fulfill (5.34) becomes

$$2^k k^{m-1} \simeq N_k \lesssim \frac{n}{\log(n)}. \tag{5.35}$$

The overall error

Let (5.34) be fulfilled. Then, theorem 5.15 provides the upper bound

$$E \lesssim \inf_{f \in V_k} \|f - g\|^2_{L_2(T;\mathbb{C}^d)} + n^{-\sigma} \overset{(3.51)}{\lesssim} d 2^{-2sk} \|g\|^2_{\bar{H}^s_{\mathrm{mix}}(T;\mathbb{C}^d)} + n^{-\sigma} \lesssim 2^{-2sk} + n^{-\sigma} \tag{5.36}$$

on the error $E := \mathbb{E}_{\rho_T^n}\left[\mathcal{E}\left(\tau_r\left(f_{\mathcal{Z}_n,V_k}\right)\right) - \mathcal{E}\left(g\right)\right]$. Here, the $\lesssim$ constant in the final estimate depends on m, d, σ, r, s and $\|g\|_{\bar{H}^s_{\mathrm{mix}}(T;\mathbb{C}^d)}$.

Balancing the error terms

Let k be the largest natural number such that (5.34) still holds. In this case, we have

$$2^k k^{m-1} \simeq N_k \simeq \frac{n}{\log(n)},$$

which also gives $k \lesssim \log(n)$ for $n > 1$. Therefore, substituting this scaling into (5.36), we obtain

$$\begin{aligned} E &\lesssim 2^{-2sk} + n^{-\sigma} \simeq \left(2^k k^{m-1}\right)^{-2s} k^{2s(m-1)} + n^{-\sigma} \lesssim \left(\frac{\log(n)}{n}\right)^{2s} \log(n)^{2s(m-1)} + n^{-\sigma} \\ &\simeq n^{-2s} \log(n)^{2sm} + n^{-\sigma}. \end{aligned}$$

By choosing $\sigma = 2s$, the overall error bound reads

$$E = \mathcal{O}\left(n^{-2s} \log(n)^{2sm}\right).$$

5.5.5 Overview

A summary of the results which we derived in the previous subsections can be found in table 5.2. The columns there have to be understood in the following way:

- The sufficient condition on the coupling between N_k and n guarantees the stability of solving the unregularized, noiseless regression problem (D) with probability at least $1 - 2n^{-\sigma}$.
- The balanced (optimal) scaling between n and N_k together with the choice of σ in the next column optimizes the error bound from theorem 5.15 such that both summands are approximately equal up to logarithms.
- The balanced rate in the last column of table 5.2 resembles the convergence rate of the error $\mathbb{E}_{\rho_T^n}\left[\mathcal{E}\left(\tau_r\left(f_{\mathcal{Z}_n,V_k}\right)\right) - \mathcal{E}\left(g\right)\right]$ with respect to n for the optimal scaling between n and k and the optimal σ.

Note that the balancing $N_k \simeq \frac{n}{\log(n)}$ is the same for all four cases which we considered.

(a) Piecewise linear prewavelets on full grids and and sparse grids, $0 < s \leq 2$

V_k	suff. cond.	balanced n	balanced σ	balanced rate
$\mathcal{V}_k^{\text{full},d}$	$4^m N_k \leq \Upsilon_\sigma \frac{n}{\log(n)}$	$N_k \simeq \frac{n}{\log(n)}$	$\sigma = \frac{2s}{m}$	$n^{-\frac{2s}{m}} \log(n)^{\frac{2s}{m}}$
$\mathcal{V}_k^{\text{sparse},d}$	$2\left(\frac{72}{25}\right)^m N_k \leq \Upsilon_\sigma \frac{n}{\log(n)}$	$N_k \simeq \frac{n}{\log(n)}$	$\sigma = 2s$	$n^{-2s} \log(n)^{(2s+1)m-1}$

(b) Fourier polynomials on full grids and hyperbolic crosses, $s \geq 0$

V_k	suff. cond.	balanced n	balanced σ	balanced rate
$\mathcal{T}_k^{\text{full},d}$	$N_k \leq \Upsilon_\sigma \frac{n}{\log(n)}$	$N_k \simeq \frac{n}{\log(n)}$	$\sigma = \frac{2s}{m}$	$n^{-\frac{2s}{m}} \log(n)^{\frac{2s}{m}}$
$\mathcal{T}_k^{\text{hyp},d}$	$N_k \leq \Upsilon_\sigma \frac{n}{\log(n)}$	$N_k \simeq \frac{n}{\log(n)}$	$\sigma = 2s$	$n^{-2s} \log(n)^{2sm}$

Table 5.2: Results for noiseless regression (D) over V_k for full grids (H^1 regularization, $g \in H^s$, $N_k \simeq 2^{km}$) and sparse grids/hyperbolic crosses (H^1_{mix} regularization, $g \in H^s_{\text{mix}}$, $N_k \simeq 2^k k^{m-1}$). We provide a sufficient condition on the scaling between N_k and n such that the problem is stable with probability at least $1 - 2n^{-\sigma}$. Here, we used $\Upsilon_\sigma := \lambda_{\min}(M)|\log(c_{\frac{1}{2}})|(1+\sigma)^{-1}$. Furthermore, we present the relation between n and N_k in the balanced case as well as the appropriate choice of σ. The last column of the tables contains the balanced convergence rate with respect to n.

The curse of dimensionality

We observed that - due to condition (5.11) on $K(N_k)$ - the coupling $N_k \lesssim \frac{n}{\log(n)}$ is mandatory for all our examples to ensure stability and convergence of the noiseless regression problem with high probability. This condition directly translates to the presence of the curse of dimensionality with respect to the necessary number of samples n in the full grid case since $N_k \simeq 2^{km}$ here. Thus, $2^{km} \lesssim \frac{n}{\log(n)}$ has to hold and n grows exponentially. Furthermore, the curse of dimensionality is also visible in the balanced convergence rate whose leading term is $n^{-\frac{2s}{m}}$.

For sparse grids and hyperbolic crosses, the scaling of the sufficient condition for stability and convergence is $2^k k^{m-1} \simeq N_k \lesssim \frac{n}{\log(n)}$. Therefore, the curse of dimensionality appears only in a mild form with respect to the level k.

Oversampling in the optimal/balanced case

In the balanced case, we have $N_k \simeq \frac{n}{\log(n)}$. Ignoring the logarithm for a moment, this means that only $n = cN_k$ samples are needed for some constant $c > 0$. For Fourier polynomials for instance, we have $c = \Upsilon_\sigma^{-1}$, see table 5.2. Since $\lambda_{\min}(M) = 1$ in this case, c is determined by the size of σ, which is $\frac{2s}{m}$ for full grids and $2s$ for hyperbolic crosses. Therefore, the oversampling factor for hyperbolic crosses is larger than for full grids if $m > 1$. However, when taking the scaling $N_k \simeq 2^{mk}$ or $N_k \simeq 2^k k^{m-1}$, respectively,

into account, we see that the absolute number of samples which is needed is much larger in the full grid case.

Convergence with respect to n

In contrast to the regularized, more general regression problem (B) over $V_{k,b}$, which we considered in section 4.5, we can achieve (expected) error convergence rates which are faster than n^{-1} for unregularized, noiseless function regression (D) over V_k. To this end, let us first consider prewavelets on a full grid. Here, we can achieve a convergence rate of $n^{-\frac{2s}{m}} \log(n)^{\frac{2s}{m}}$. In this case, the expected rate of the overall error decay is better than n^{-1} if $m < 2s$. Since the smoothness index s is limited from above by 2 for a prewavelet discretization, the best possible convergence rate would be $n^{-\frac{4}{m}}$ up to logarithms in the case $s = 2$ and $\sigma = \frac{4}{m}$. Here, the curse of dimensionality appears again and we see that even the best convergence rate in the noiseless regression case cannot beat n^{-1} if the dimension is too large, i.e. $m \geq 4$, in the full grid case.

For sparse grids, however, we see from table 5.2 that the factor m is no longer present in the main term of the balanced rate $n^{-2s} \log(n)^{(2s+1)m-1} = n^{-4} \log(n)^{5m-1}$ for $s = 2$ and $\sigma = 4$.

For Fourier polynomials on full grids or hyperbolic crosses, the situation is analogously to the prewavelet case. Note that, this time, s can be arbitrarily large and thus, for $s \to \infty$, we are able to observe super-algebraic convergence rates in the balanced case, i.e. the error decays faster than any inverse polynomial n^{-p} with $p \in (0, \infty)$. However, this only holds if $\sigma \to \infty$. Since our analysis is only valid for fixed s and σ, we cannot directly deduce what happens if $s \to \infty$ and how the balanced relation between n and N_k has to look like in this case.

5.5.6 Relation to other results

We now relate our findings to other works in the research area of penalized and unpenalized regression.

Convergence and stability results for arbitrary orthonormal bases

Naturally, we have to compare our results to [21] since the basic ideas for our proofs are provided there and it can be seen as the foundation for analyzing the higher order convergence rates with respect to n for noiseless least-squares regression in a very general setting. Their theorems 1 and 2 can be interpreted as special cases of our theorems 5.11 and 5.15 for $d = 1$, an orthonormal basis $\nu_1, \ldots, \nu_{N_k}$ of V_k and the unregularized case $\mu = 0$. In this specific situation, our results read the same as the ones in [21]. Furthermore, the authors also consider a Fourier polynomial example in the univariate case and show that $t_i = -\pi + 2\pi \frac{j}{n}$ provides a deterministic point distribution for $j = 1, \ldots, n$ such that stability of the corresponding least-squares algorithm is guaranteed if $n \geq N_k$, i.e. for

this deterministic sample, the right hand side in the stability condition (5.31) becomes n. Note that the authors also provide the error bound

$$\mathbb{E}_{\rho_T^n}\left[\mathcal{E}\left(\tau_r\left(f_{\mathcal{Z}_n,V_k}\right)\right)-\mathcal{E}\left(g\right)\right] \lesssim \inf_{f\in V_k}\|f-g\|^2_{L_{2,\rho_T}(T)}+\frac{N_k}{n}$$

for unregularized regression in the presence of additive noise, which (asymptotically) coincides with the bounds we obtained in chapter 4 up to logarithms, cf. theorems 4.18, 4.24 and the discussion in subsection 4.5.6.

Least-squares regression with global polynomials

Based on [61], the approach from [21] has been applied to the case of global polynomial bases in total degree spaces and hyperbolic cross spaces in [19]. There, the authors also show the validity of their results in the Hilbert space-valued case by exploiting the tensor product identity $L_{2,\rho_T}(T;X) = X \otimes L_{2,\rho_T}(T)$ and using the scalar-valued results. They provide bounds on $K(N_k)$ for several types of tensorized Jacobi-polynomials, such as Legendre polynomials ($K(N_k) \lesssim N_k^2$), even for infinite-dimensional T and E. Furthermore, when changing the measure ρ_T to the Chebyshev measure with density $\frac{1}{\sqrt{1-t^2}}$ in each direction, they show that Chebyshev polynomials are able to achieve

$$K(N_k) \lesssim N_k^{\frac{\log(3)}{\log(2)}}$$

in the infinite-dimensional case $m = \infty$ and

$$K(N_k) \leq 2^m N_k$$

in the finite-dimensional case $m < \infty$. In the finite-dimensional case, this is similar to the conditions $K(N_k) \leq 4^m N_k$ and $K(N_k) \leq 2\left(\frac{72}{25}\right)^m N_k$ we proved for prewavelets on full grids and sparse grids with respect to the Lebesgue measure $\rho_T = \lambda_T$, cf. subsections 5.5.1 and 5.5.2. Note however, that the authors of [19] usually deal with densely populated sample matrices G because of their global bases, while our spline basis functions have only local support and lead to sample matrices with sparse ("finger-like") structure, see also [11, 31]. A detailed analysis of probabilistic error bounds for function regression also for non-centered additive noise models can be found in [60]. In [59], the stability and convergence properties of least-squares regression on global polynomial search spaces with deterministic samples $\mathbf{t}_i$ for $i = 1, \ldots, n$ from low discrepancy point sets have been studied.

Compressive sensing and the restricted isometry property

Considering a compressive sensing approach to the regression problem, the aim is to find sparse solutions. The corresponding constrained minimization problem to obtain the

function $f = \sum_{i=1}^{N_k} \beta_i \nu_i$ can then be written as[5]

$$\min_{\vec{\beta} \in \mathbb{R}^{N_k}} \|\beta\|_{\ell_1} \text{ such that } G\vec{\beta} = B(\vec{\mathbf{x}})$$

with G, B and $\vec{\mathbf{x}}$ from propositions 5.4 and 5.5. Here, the minimization of the ℓ_1 norm of the coefficients guarantees the sparsity of the solution, i.e. most entries of $\vec{\beta}$ are 0, see e.g. [25, 32].

Note that, for our examples above, the regularization term in the least-squares approach can always be written as a sum of weighted L_2 norms and, ultimately, as a weighted ℓ_2 norm of the coefficients, see e.g. (3.37) and (3.38) for the prewavelet spline spaces and Sobolev norm regularizations. Therefore, the least-squares regression problem (C) is similar to a dual formulation of the above compressive sensing problem in this case. The main difference is now the appearance of the ℓ_1 norm for compressive sensing and the weighted ℓ_2 norm in our case.

In the following, we consider scalar-valued functions, i.e. $d = 1$. A natural question which arises in the context of compressive sensing is if $G\vec{\beta} = B(\vec{\mathbf{x}})$ is solvable for an s-sparse vector $\vec{\beta}$, i.e. a vector with at most s non-zero entries. To this end, one usually aims to establish the so-called *restricted isometry property (RIP)*. The RIP basically says that G is almost an isometry for every s-sparse vector, i.e. there exists $0 < \delta < 1$ such that

$$(1-\delta)\|\vec{\beta}\|_2^2 \leq \|G\vec{\beta}\|_2^2 \leq (1+\delta)\|\vec{\beta}\|_2^2 \quad \forall\, s\text{-sparse } \beta \in \mathbb{R}^{N_k}.$$

Thus, the best s-sparse solution which minimizes $\|G\vec{\beta} - B(\vec{\mathbf{x}})\|_2$ is computable if the RIP holds. To establish the RIP with high probability, the parameter s has to be appropriately coupled to the number of samples n. To this end, it has recently been shown in [20] that for Chebyshev and Legendre polynomials - and the corresponding measures ρ_T - the number

$$\tilde{K}(s) := \sup_{\substack{\Lambda \subset \mathcal{I}_{\text{hyp}}(\log(s)) \\ |\Lambda| = s \\ \Lambda \text{ lower}}} \left\| \sum_{\mathbf{l} \in \Lambda} |\nu_{\mathbf{l}}|^2 \right\|_{L_{\infty,\rho_T}(T)}$$

plays an important role for such a coupling. Here, $\nu_{\mathbf{l}}$ is the corresponding polynomial of degree $\mathbf{l}$, $\mathcal{I}_{\text{hyp}}(\log(s))$ is the index set of a hyperbolic cross of level $\log(s)$ and we call a set $\Lambda \subset \mathbb{N}^m$ *lower* if $\mathbf{l} \in \Lambda$ implies $\mathbf{k} \in \Lambda$ for all $\mathbf{k} \leq \mathbf{l}$. Note that $\tilde{K}(s)$ is nothing else but the supremum of $K(N_k)$ over all collections of functions which result from lower subsets of $\mathcal{I}_{\text{hyp}}(\log(s))$ of size s. The authors of [20] have shown that there exists a constant $C > 0$, independent of s and m, such that

$$C\tilde{K}(s)\log(s)^2(\log(s)+m) \leq n$$

[5] Note that it often is only required that the equations are solved up to a certain error $\varepsilon > 0$.

suffices to establish an RIP for lower index subsets of $\mathcal{I}_{\text{hyp}}(\log(s))$ with s elements. This is directly related to the stability condition

$$K(N_k) \lesssim \frac{n}{\log(n)},$$

for least-squares regression, see also (5.11). Apart from the logarithmic factors, the difference between these two inequalities is the substitution of $K(N_k)$ by $\tilde{K}(s)$ for the compressive sensing case. Thus, the stability condition for least-squares regression becomes a sufficient condition on the sparsity s to establish an RIP with high probability. A similar analysis as in [20], but for an arbitrarily weighted ℓ_1 norm minimization, can be found in [2].

5.6 Summary

We now provide a brief summary of our results to conclude this chapter:

- We have shown the equivalence between the primal problem of solving (B) over $\mathcal{H}_b$ and the dual problem of solving

$$\text{Find} \quad \arg\min_{f \in \mathcal{H}} \mathcal{E}_{\mathcal{Z}_n}(f) + \mu \|f\|_{\mathcal{H}}^2 \tag{C}$$

  for a certain Lagrange parameter $\mu \geq 0$. A solution of the latter problem can be computed with the help of the representer theorem 5.3 if $\mathcal{H}$ is an RKHS. For finite-dimensional search spaces, we can solve the system

$$(G + \mu C)\vec{\alpha} = B(\vec{\mathbf{x}})$$

  of linear equations instead and obtain the solution $f_{\mathcal{Z}_n, V_k, \mu} = \sum_{j=1}^{N_k} \alpha_j \nu_j$. Here, the size of the system is determined by N_k instead of n, where the latter would be the case if the representer theorem was applied.

- We proved that solving the above system is (uniformly) stable for all $k \in \mathbb{N}$ with probability $1 - 2n^{-\sigma}$ if

$$K(N_k) \leq \frac{\lambda_{\min}(M) |\log(c_{\frac{1}{2}})|}{(1+\sigma)} \cdot \frac{n}{\log(n)}$$

  holds and if

  - $\nu_1, \ldots, \nu_{N_k}$ is a Riesz basis or
  - $\mu \cdot \lambda_{\min}(C)$ is bounded from below independently of k.

- Similarly to the bias/sampling error decomposition in chapter 4, we observed that the expected overall error

$$E := \mathbb{E}_{\rho_T^n}\left[\mathcal{E}\left(\tau_r\left(f_{\mathcal{Z}_n,V_k}\right)\right) - \mathcal{E}\left(g\right)\right]$$

for an r-truncated solution of the unregularized, noiseless function regression problem

$$\text{Find} \quad f_{\mathcal{Z}_n,V_k} = \arg\min_{f\in V_k} \frac{1}{n}\sum_{i=1}^{n} \|f(\mathbf{t}_i) - g(\mathbf{t}_i)\|_E^2 \tag{D}$$

can be bounded by

$$E \leq \left(1 + \frac{4\lambda_{\max}(M)|\log(c_{\frac{1}{2}})|}{(1+\sigma)\lambda_{\min}(M)\log(n)}\right) \inf_{f\in V_k} \|f-g\|^2_{L_{2,\rho_T}(T;E)} + 8r^2 n^{-\sigma}$$

if the above condition on $K(N_k)$ holds.

- We investigated the stability conditions and the convergence rates of unregularized, noiseless function regression on sparse grids and hyperbolic crosses. Here, we observed the considerably reduced effect of the curse of dimensionality in contrast to the full grid methods and showed that the convergence can be faster than n^{-1}.

6 Numerical Experiments

After deriving our theoretical statements in the previous chapters, we now underpin our findings by several numerical results. As we already mentioned in section 5.1, the computation of a solution to the constrained problem (B) over the set $V_{k,b}$ is very involved. Therefore, we restrict[1] ourselves to the dual problem (C) and solve (5.6).

For the prewavelet spline case, we used the C++ *sparselib* code developed in [29, 30], which we maintained and enhanced during the last years. This code features efficient traversal algorithms for prewavelet basis functions as well as fast matrix-vector multiplication routines based on the *unidirectional principle*, see also [13, 91]. To solve the system of linear equations we employ a conjugate gradient algorithm which stops if the ℓ_2 norm of the relative residual is smaller than 10^{-13} or if the number of iterations reaches $\max(10^4, N_k)$. The latter criterion is only employed to avoid problems in the case of very ill-conditioned systems, e.g. in the case of small (or no) regularization and $N_k > n$. Due to the fast matrix-vector multiplication routines, the computational complexity of each CG-step is only $\mathcal{O}\left(N_k + nk^{m-1}\right)$, see also [10, 31], and the overall number of CG steps is small in most cases, i.e. if $n > N_k$ or if the regularization parameter is large enough.

For the Fourier polynomial experiments, we wrote a python program based on the *numpy* library. Due to the densely populated G matrix, we employ an LU decomposition with partial pivoting to solve the least-squares problem on full grids and hyperbolic crosses. The computational costs for the assemblation and the computation of the solution scale like $\mathcal{O}\left(N_k^3 + n \cdot N_k^2\right)$, see [70].

As in the example sections 4.5 and 5.5, we employ an $H^1/\bar{H}^1$ regularization in the full grid case and an $H^1_{\text{mix}}/\bar{H}^1_{\text{mix}}$ regularization in the sparse grid/hyperbolic cross case. Here, we use the norm equivalences (3.37), (3.38), (3.46) and (3.47) to compute the regularization matrix, i.e. we approximate C by the corresponding diagonal weight matrix instead of taking the true regularization matrix.

We start with a thorough convergence study for a noisy scalar-valued function regression problem in two dimensions in section 6.1. Due to the noise, we cannot rely on the convergence results of section 5.5 but can only hope to achieve the convergence rates predicted in section 4.5. Next, we study the noiseless case in section 6.2. In section 6.3, we have a look at a problem which employs a non-smooth solution. We briefly introduce a dimension-adaptive sparse grid regression algorithm, which is able to improve the convergence with respect to N_k in this case. Finally, after considering the aforementioned artificial problems, we deal with real world examples in section 6.4.

[1] Note, however, that we compared some of our results for small problem instances to the results of a constrained minimization algorithm for (B) and they did not differ significantly.

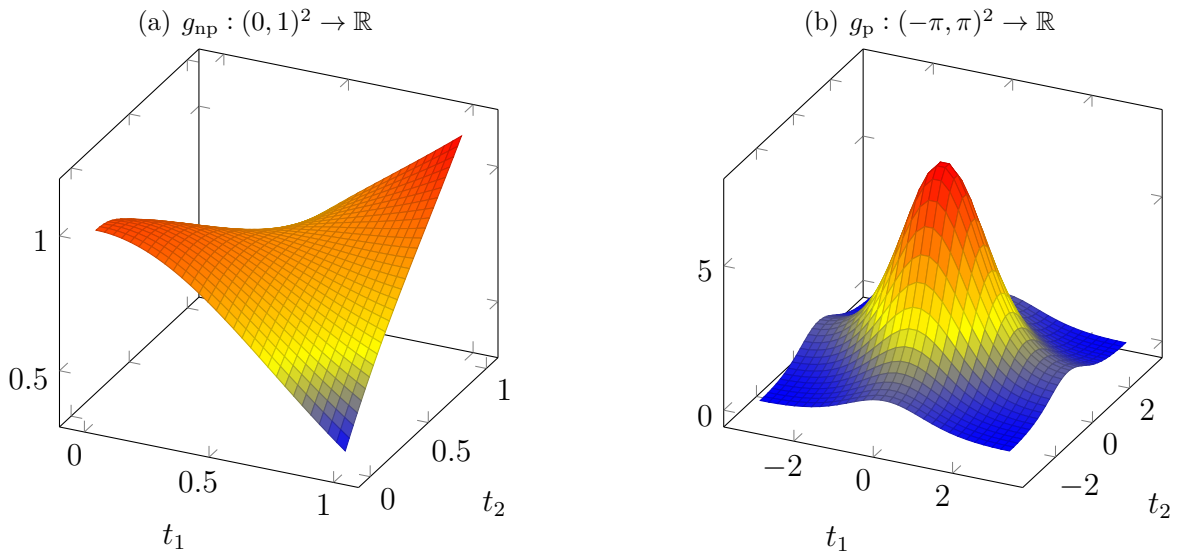

Fig. 6.1: The example functions g_{np} and g_{p}

6.1 Convergence analysis for regularized regression of noisy data

We first have a look at two scalar-valued function regression problems which admit a smooth solution. In the non-periodic case, we consider

$$g_{\mathrm{np}}(\mathbf{t}) = \exp\left(-\|\mathbf{t}\|_2^2\right) + \prod_{i=1}^{m} t_i$$

on $T = (0,1)^m$ with dimension $m = 2$. This smooth function served as a benchmark example for sparse grid regression already in [35]. Obviously, $g_{\mathrm{np}} \in H^2(T)$ and $g_{\mathrm{np}} \in H^2_{\mathrm{mix}}(T)$. In the periodic setting, we take the smooth test function

$$g_{\mathrm{p}}(\mathbf{t}) = \exp\left(\sum_{i=1}^{m} \cos\left(t_i\right)\right)$$

on $T = (-\pi,\pi)^m$ with dimension $m = 2$, which fulfills $g_{\mathrm{p}} \in \bar{H}^s(T)$ and $g_{\mathrm{p}} \in \bar{H}^s_{\mathrm{mix}}(T)$ for every $s > 0$. The shape of both test functions is depicted in figure 6.1.

We choose normally distributed additive noise with variance 0.01. Therefore, to sample the conditional measure $\rho(\mathbf{x}|\mathbf{t})$, $\mathbf{x}$ is drawn according to the normal distribution $N(g_*(\mathbf{t}), 0.01)$, where $* = \mathrm{np}$ in the non-periodic case and $* = \mathrm{p}$ in the periodic case. We directly observe that $f_\rho = g_*$ in this setting. The distribution on T is chosen to be uniformly in both cases, i.e. $\rho_T = \lambda_T$ for $T = (0,1)^2$ and $\rho_T = \frac{1}{(2\pi)^2}\lambda_T$ for $T = (-\pi,\pi)^2$. Since we are dealing with noisy regression, we cannot apply theorem 5.15 and have to

rely on our convergence results from chapter 4. Note that, technically, our setup does not fulfill the prerequisites of chapter 4 since r from (4.5) is infinite for Gaussian noise. However, because of the fast decay of the tails of the Gauß density, the error of truncating it at a certain, large enough value is negligibly small. Therefore, we tacitly assume that we can apply our theorems anyhow. As we will observe in the following, the numerical results match the outcome of our theoretical analysis in section 4.5 for the choice $M_\psi \simeq 1$. This shows again that it is often justified to assume that the M-boundedness constant M_ψ can be chosen independently of the regularization parameter, even though there is no direct theoretical justification.

6.1.1 Non-periodic regression on full grids and sparse grids

We start with a study of the regression error

$$\text{Err} := \mathcal{E}\left(f_{\mathcal{Z}_n,V_k,\mu}\right) - \mathcal{E}\left(g_{\text{np}}\right) = \|f_{\mathcal{Z}_n,V_k,\mu} - g_{\text{np}}\|^2_{L_2(T)}$$

in the full grid space and the sparse grid space for noisy samples of g_{np}. To this end, we computed the solutions $f_{\mathcal{Z}_n,V_k,\mu}$ to the dual problem for $V_k = \mathcal{V}_k^{*,1}$ with $* \in \{\text{full}, \text{sparse}\}$ for various choices of the sample size n and the grid level k. To evaluate the error, we interpolated g_{np} on a full grid of level 11 and used the mass matrix on this level to compute the squared L_2 norm. For each parameter tuple (k, n), we ran 10 individual computations for different, random input data points and calculated the arithmetic mean AvErr of Err over these 10 runs. We showed in subsections 4.5.1 and 4.5.2 that $b \simeq 2^k$ is an appropriate choice for the norm bound in the primal problem, see also the overview in subsection 4.5.5. However, it is not directly clear how this can be transferred to the choice of the regularization parameter μ. Therefore, we studied three different cases: the constant coupling (CC) $\mu = 10^{-3}$, the linear coupling (LC) $\mu = 10^{-3} \cdot 2^{-k} \simeq 10^{-3} b^{-1}$ and the quadratic coupling (QC) $\mu = 10^{-3} \cdot 2^{-2k} \simeq 10^{-3} b^{-2}$.

The overall error in dependence on k and n

The plots for the average error for the full grid can be found in figure 6.2 and the average error for the sparse grid is depicted in figure 6.3. We see that the general behavior is the same in both cases. We first discuss the error behavior for fixed n and varying k, i.e. the plots in the top row of the figures. First of all, we observe that the CC $\mu = 10^{-3}$ is too restrictive, i.e. we see no convergence of the overall error after an initial reduction to 10^{-5}. For the LC $\mu = 10^{-3} 2^{-k}$, however, the results are more promising. In plot (b) of figures 6.2 and 6.3, we indeed observe the decay rates 2^{-4k} and $2^{-4k} \cdot k$ in the full and the sparse grid case, respectively. Because of the smoothness of g_{np}, these rates for the discretization error are also predicted by (4.46) and (4.52) with $s = 2$. Note, however, that the error increases again for $k \geq 5$ since the sampling error now dominates in the overall error decomposition and the noisy data points are overfitted more and more for

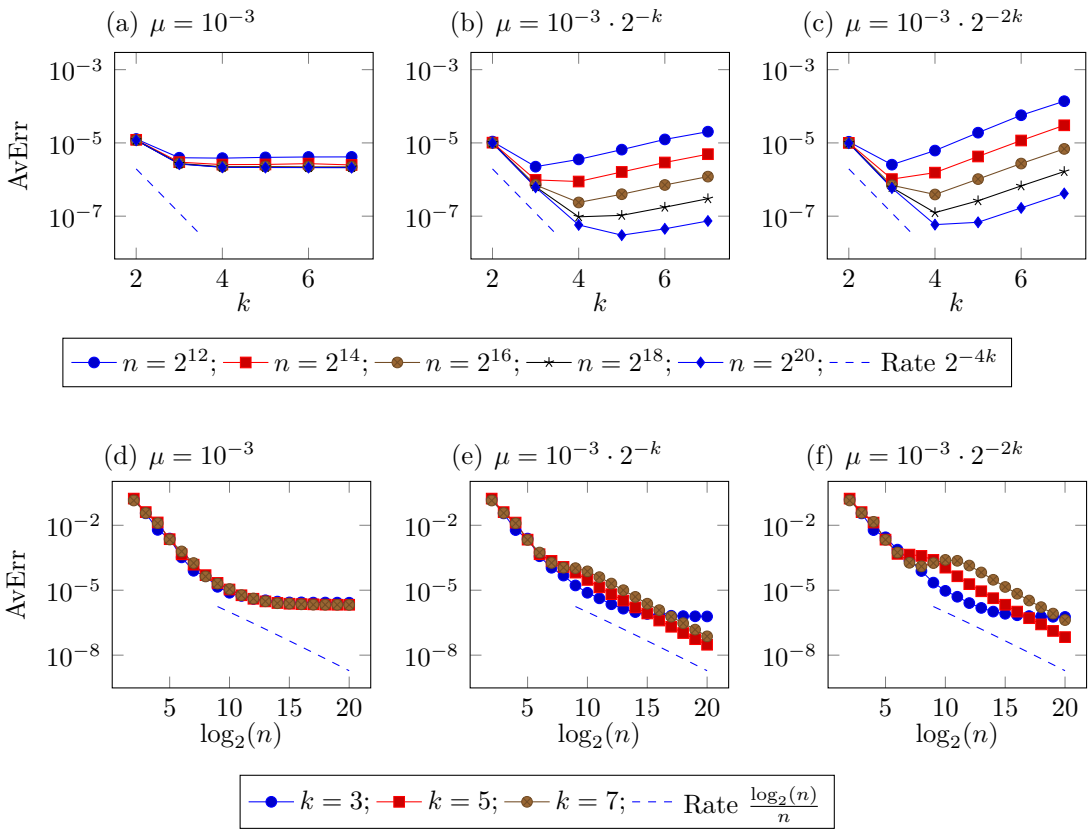

Fig. 6.2: Full grid $V_k = \mathcal{V}_k^{\text{full},1}$: The average AvErr of the overall error $\text{Err} = \mathcal{E}\left(f_{\mathcal{Z}_n, V_k, \mu}\right) - \mathcal{E}\left(g_{\text{np}}\right)$ over 10 independent draws of input data $\mathcal{Z}_n$ is plotted versus the level k for different choices of n (top) and versus the number of data n for different choices of k (bottom).

increasing grid level. Although, the results for the QC $\mu = 10^{-3}2^{-2k}$ are quite similar, subfigure (c) shows that the overall error increases much faster for $k \geq 5$ than in the LC case since the overfitting is more severe in the QC case.

Now, we consider the behavior of the overall error for fixed k and varying n (bottom row of figures 6.2 and 6.3). Here, we have to discern two regimes. For small n, we observe a convergence rate of approximately n^{-2} and for larger n the rate becomes $\frac{\log_2(n)}{n}$. The latter matches our theoretical bounds on the sampling error, see also (4.46) and (4.52). The faster convergence in the beginning is due to the fact that we are still above the noise level there. Note the plateau in the QC case (subfigure (f)) for $k = 7$ (and also $k = 9$ for the sparse grid) around $\log_2(n) = 10$. This is again due to the severe overfitting of the noisy data points. Furthermore, the error for $k = 3$ stagnates for large n. This resembles the fact that, here, the discretization error dominates and the level k needs to be refined in order to achieve a smaller overall error. In summary, we observe that it is not reasonable to go beyond $k = 5$ even for $n = 2^{20}$ data points. We will see that this is

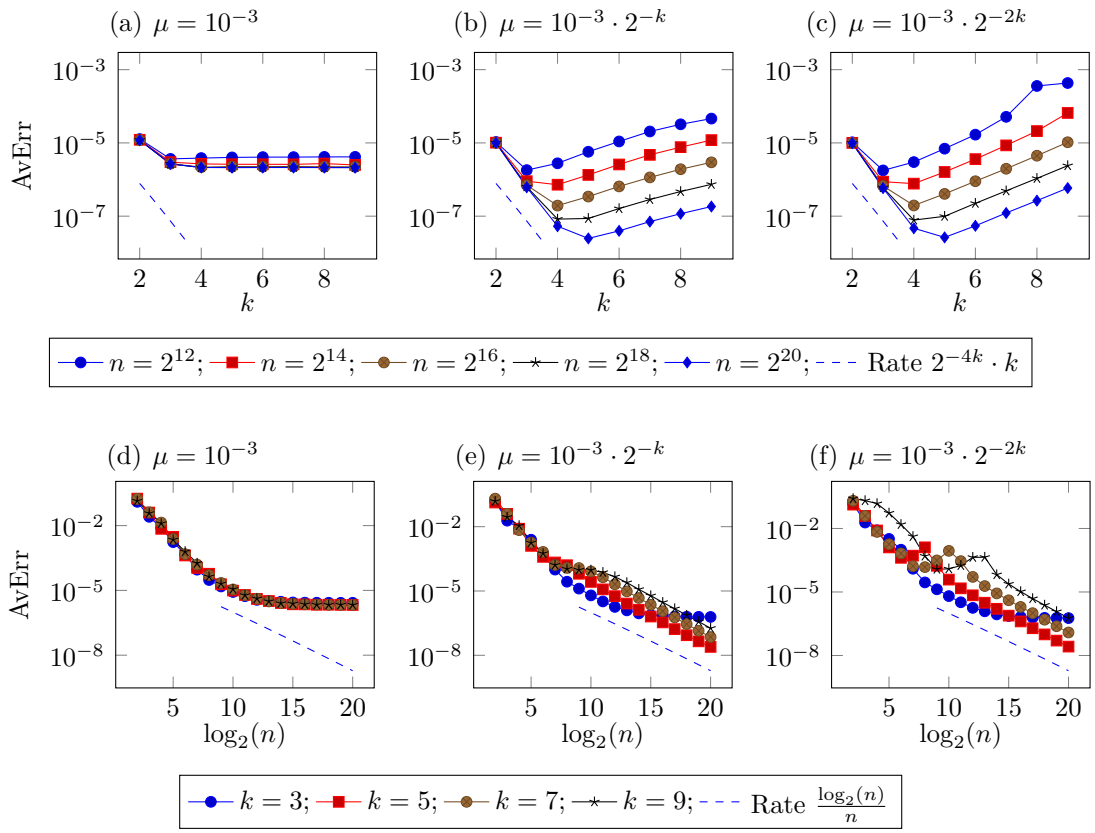

Fig. 6.3: Sparse grid $V_k = \mathcal{V}_k^{\text{sparse},1}$: The average AvErr of the overall error $\text{Err} = \mathcal{E}(f_{\mathcal{Z}_n,V_k,\mu}) - \mathcal{E}(g_{\text{np}})$ over 10 independent draws of input data $\mathcal{Z}_n$ is plotted versus the level k for different choices of n (top) and versus the number of data n for different choices of k (bottom).

also reflected by the optimal balancing between k and n, which we investigate later on.

The overall error in dependence on N_k

Although our results suggest that the full grid method and the sparse grid method behave essentially in the same way, this is of course only true with respect to the discretization level k and not with respect to the basis size N_k. To clarify this, we depicted the behavior of the overall error with respect to N_k for $n = 2^{20}$ in figure 6.4. There, we also depicted the error when considering the analogous convergence study in $m = 3$ dimensions to underpin the fact that the spread between the full grid and the sparse grid results becomes even more obvious in higher dimensions. We observe that, for sparse grids, AvErr is at its minimum for a significantly smaller value of N_k than for full grids. This means that we need less degrees of freedom for sparse grids to reach the margin where the sampling error begins to prevail.

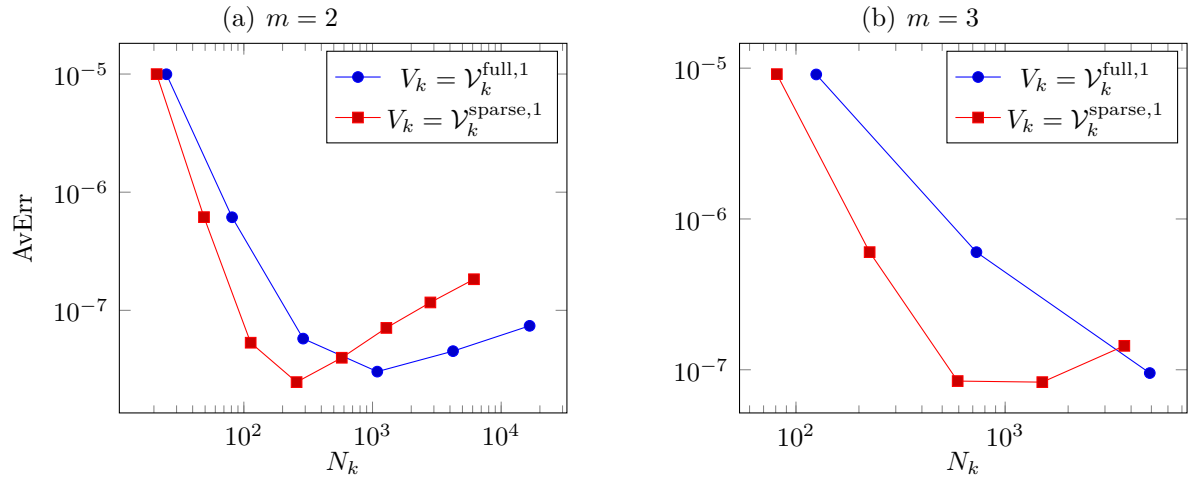

Fig. 6.4: The average overall error AvErr in dependence on the basis size N_k for full grids and sparse grids with $n = 2^{20}, \mu = 10^{-3}2^{-k}$. We plotted the two-variate case $m = 2$ (left) and the three-variate case $m = 3$ (right).

Stability

Besides the error decay, we are also interested in the stability when solving the linear system. Since theorem 5.11 can also be applied in the noisy function regression case, we expect that the condition number of our system matrix $G + \mu C$ is small with high probability if $K(N_k) \lesssim \frac{n}{\log(n)}$, see (5.11). Because of our considerations in subsections 5.5.1 and 5.5.2, we know that $K(N_k) \simeq N_k$ and the above inequality becomes $N_k \lesssim \frac{n}{\log(n)}$. A closer look at the $K(N_k)$ condition (5.27) for sparse grids reveals that the oversampling constant c involved in $cN_k \leq \frac{n}{\log(n)}$ needs to be larger than

$$\frac{(1+\sigma)2 \cdot (2.88)^m}{\lambda_{\min}(M) \cdot |\log(c_{\frac{1}{2}})|}.$$

For $m = 2$ and for the realistic choice $\sigma \approx 1$ and $\lambda_{\min}(M) \approx 0.1$, we obtain that c needs to be larger than $3066 \approx 2^{11.6}$ for theorem 5.11 to hold. This results in the oversampling condition $2^{11.6}N_k \leq \frac{n}{\log(n)}$, which essentially is $11.6 + k \leq \log_2(n)$ if we omit double-logarithmic terms in n and N_k. A similar analysis can also be performed for the full grid case with dimension-dependent prefactor 4^m, see (5.23), instead of $2 \cdot (2.88)^m$, which leads to a comparable constant c as in the sparse grid case. However, here we have $N_k \simeq 2^{mk} = 2^{2k}$ and the oversampling condition becomes approximately $11.6 + 2k \leq \log_2(n)$.

We calculated the average condition numbers $\kappa(S) = \frac{\lambda_{\max}(S)}{\lambda_{\min}(S)}$ over the 10 computations for each parameter set (k, n) using a singular value decomposition of $S = G + \mu C$ and plotted the results in figures 6.5 and 6.6. There, we essentially discern between

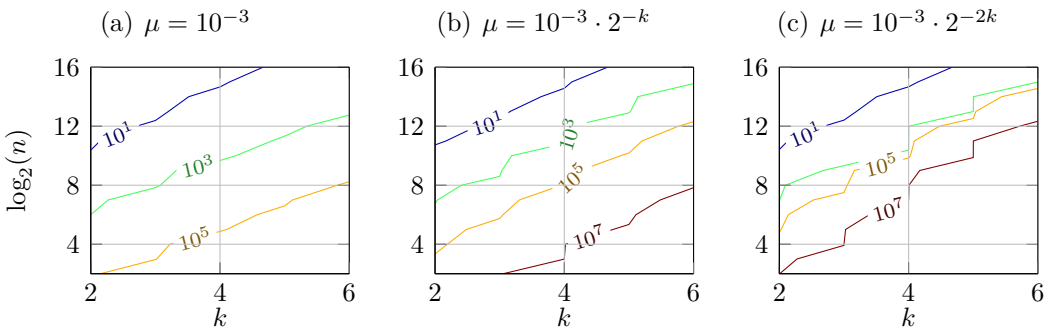

Fig. 6.5: Contour plot of the average of the condition numbers $\kappa(G + \mu C)$ over 10 independent draws of input data $\mathcal{Z}_n$ for full grids $V_k = \mathcal{V}_k^{\text{full},1}$. We depicted the contour lines for $\kappa = 10, 10^3, 10^5$ and 10^7.

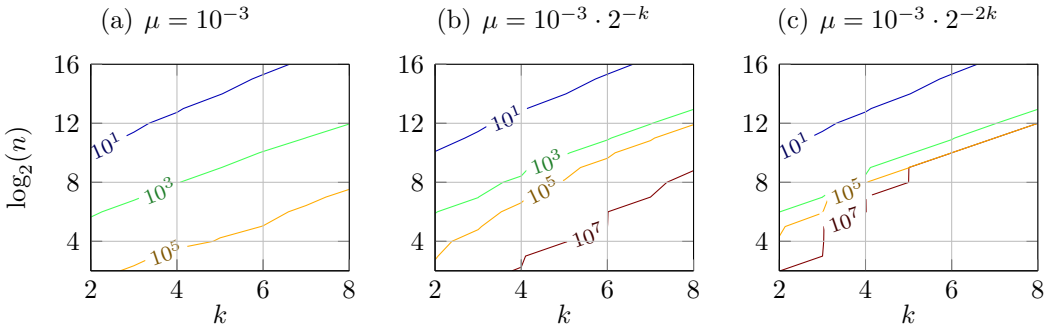

Fig. 6.6: Contour plot of the average of the condition numbers $\kappa(G + \mu C)$ over 10 independent draws of input data $\mathcal{Z}_n$ for sparse grids $V_k = \mathcal{V}_k^{\text{sparse},1}$. We depicted the contour lines for $\kappa = 10, 10^3, 10^5$ and 10^7.

well-conditioned systems ($\kappa(S) < 10$), treatable systems ($10 \leq \kappa(S) < 10^7$) and ill-conditioned systems ($\kappa(S) \geq 10^7$). First, we observe that the system is always well-conditioned if our derived oversampling condition ($11.6 + 2k \leq \log_2(n)$ for full grids and $11.6 + k \leq \log_2(n)$ for sparse grids) is fulfilled. The conditions even seem to be a bit too pessimistic when considering the isoline for $\kappa(S) = 10$ in the plots. Besides, there are many more parameter tuples (k, n) for which the equation system is still treatable and employs condition numbers smaller than 10^7. For the CC case with $\mu = 10^{-3}$, we see that all plotted pairs of k and n fall into that category. In the LC case, only scenarios with (approximately) $k > \log_2(n)$ lead to ill-conditioned system matrices. However, as we have seen in our convergence studies in figures 6.2 and 6.3, and also in our analysis in section 4.5, we need to have $\log_2(n) > k$ in order to obtain convergence of the overall error, anyhow. For QC, the parameter range, where we encounter ill-conditioned systems increases to $2k > \log_2(n)$ for full grids and $\frac{3}{2}k > \log_2(n)$ for sparse grids. Furthermore, the set of parameters in the transition area between well-conditioned systems and ill-

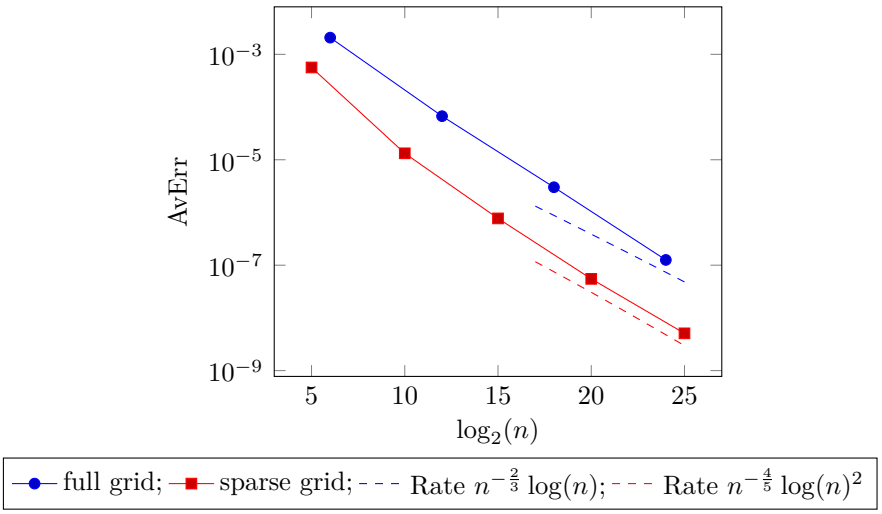

Fig. 6.7: The average overall error AvErr with $\mu = 10^{-3}2^{-k}$ after balancing the discretization error and the sampling error. For full grids, the balancing $n = 2^{6k}$ is used, whereas for sparse grids, $n = 2^{5k}$ is taken. The respective expected rates are also plotted, compare table 4.2.

conditioned ones becomes quite small.

Combining our convergence study and our stability analysis, we conclude that the linear coupling $\mu = 10^{-3}2^{-k}$ is the most promising one to achieve both a good convergence behavior of the overall error and treatable linear systems also for $\log_2(n)$ close to k.

Balanced convergence rate

Finally, we consider the convergence behavior in the case of balanced discretization error and sampling error. Since $m = s = 2$ for our example, a look at the case $M_\psi \simeq 1$ of table 4.2 (a) reveals that the appropriate scaling is $n \simeq 2^{(2s+m)k} = 2^{6k}$ for full grids and $n \simeq 2^{(2s+1)k} = 2^{5k}$ for sparse grids. The corresponding error convergence rates for this coupling can be found in figure 6.7. According to table 4.2, the upper bounds on the convergence rates are $n^{-\frac{2}{3}}\log(n)$ for full grids and $n^{-\frac{4}{5}}\log(n)^2$ for sparse grids. It can be observed that our proven rate in the balanced case seems to match our experimental results for sparse grids. However, our proven rate in the full grid case seems to be too pessimistic as the experimental rate is similar to the sparse grid rate. Possible reasons for this are, for example, that the LC for μ, although reasonable, might not correspond to $b \simeq 2^k$, or the simple fact that our proven rates are not tight and the theory does not exploit the special structure of the noisy function regression problem at hand.

It is noteworthy that, although the convergence rates with respect to n are approximately equal here, the basis size for the full grid scales like $N_k \simeq 2^{km} = 2^{2k} \simeq n^{\frac{1}{3}}$, while the basis size for the sparse grid is only $N_k \simeq 2^k k \simeq n^{\frac{1}{5}}\log(n)$.

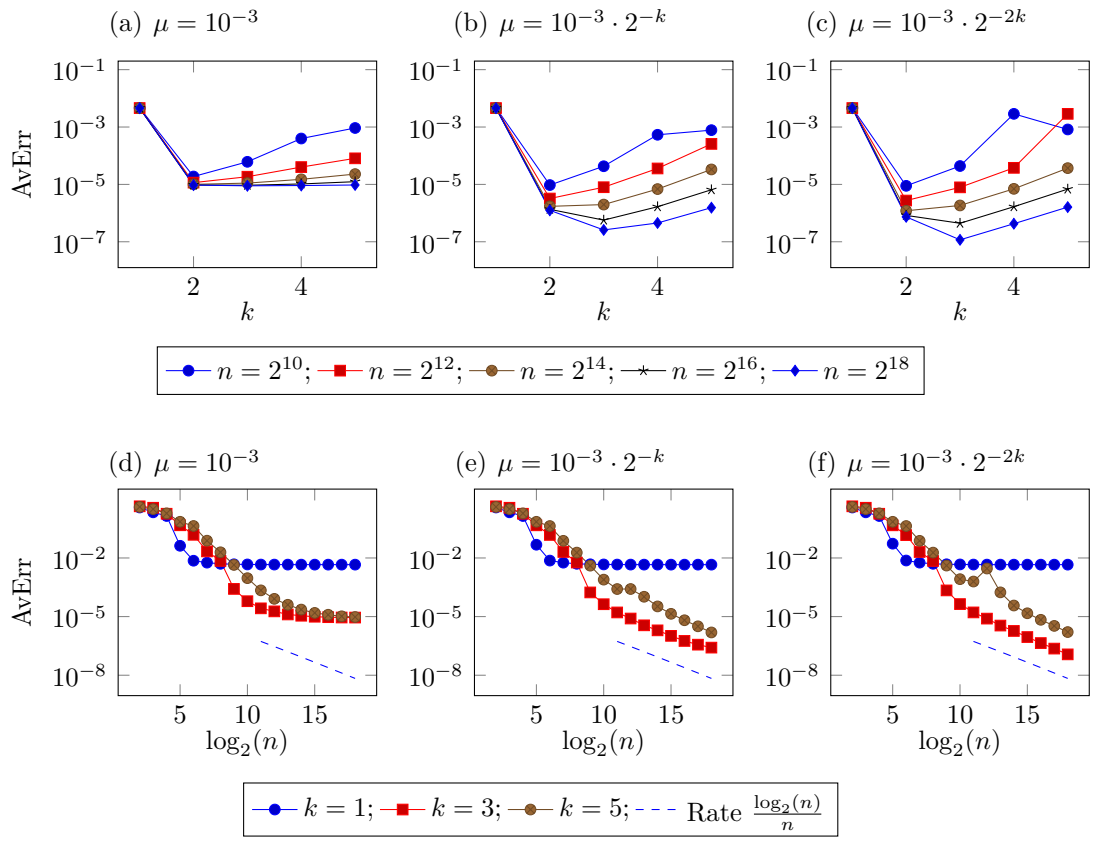

Fig. 6.8: Full grid $V_k = \mathcal{T}_k^{\text{full},1}$: The average AvErr of the overall error $\text{Err} = \mathcal{E}\left(f_{\mathcal{Z}_n,V_k,\mu}\right) - \mathcal{E}\left(g_{\text{p}}\right)$ over 10 independent draws of input data $\mathcal{Z}_n$ is plotted versus the level k for different choices of n (top) and versus the number of data n for different choices of k (bottom).

6.1.2 Periodic regression on full grids and hyperbolic crosses

We now consider the noisy function regression of g_{p}. Here, we again measure the average error AvErr over 10 random test instances of

$$\text{Err} := \|f_{\mathcal{Z}_n,V_k,\mu} - g_{\text{p}}\|_{L_2(T)}^2,$$

as in the previous subsection. However, now the sampling is done with respect to g_{p} and the solutions $f_{\mathcal{Z}_n,V_k,\mu}$ are computed for $V_k = \mathcal{T}_k^{*,1}$ with $* \in \{\text{full}, \text{hyp}\}$.

The overall error in dependence on k and n

We depicted the results for the full grid in figure 6.8 and the ones for the hyperbolic cross in figure 6.9. The qualitative behavior of the error is the same as in the previous subsection on non-periodic regression with full grids and sparse grids. We obtain that the

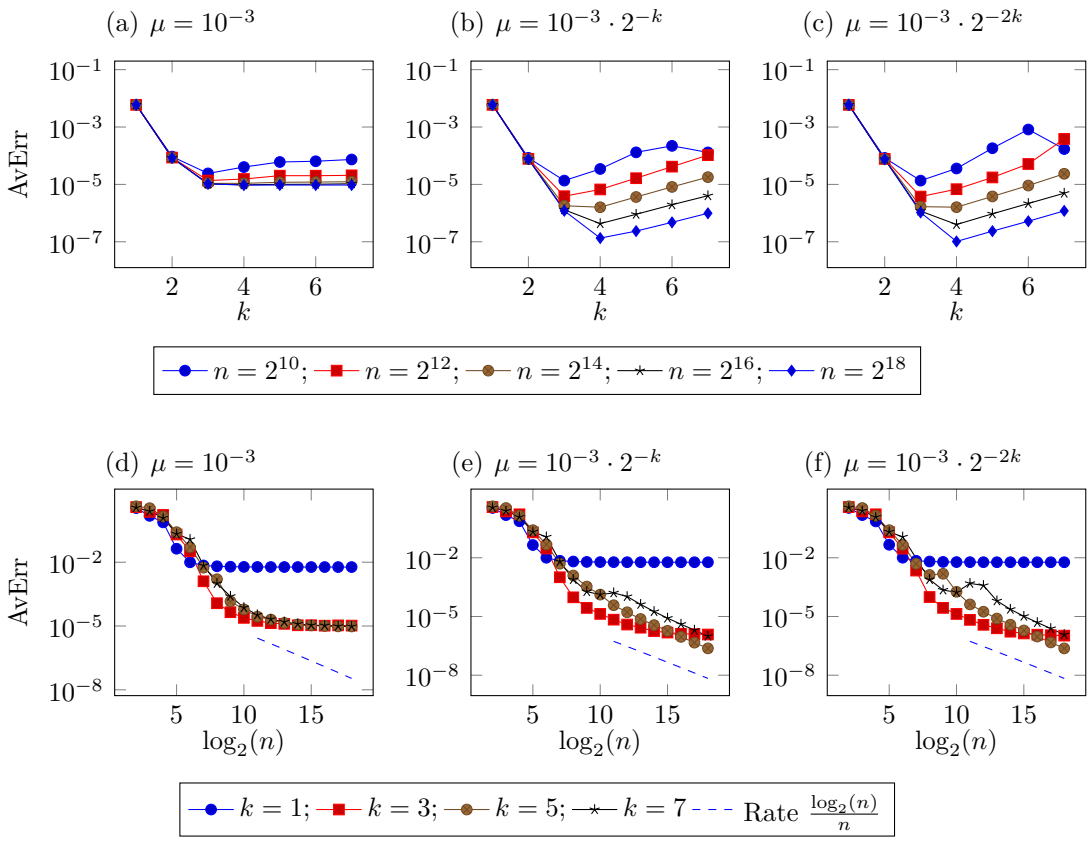

Fig. 6.9: Hyperbolic cross $V_k = \mathcal{T}_k^{\mathrm{hyp},1}$: The average AvErr of the overall error $\mathrm{Err} = \mathcal{E}\left(f_{\mathcal{Z}_n,V_k,\mu}\right) - \mathcal{E}\left(g_{\mathrm{p}}\right)$ over 10 independent draws of input data $\mathcal{Z}_n$ is plotted versus the level k for different choices of n (top) and versus the number of data n for different choices of k (bottom).

constant coupling $\mu = 10^{-3}$ leads to a too restrictive regularization and the error cannot decay after an initial drop. In the LC case $\mu = 10^{-3} \cdot 2^k$ and the QC case $\mu = 10^{-3} \cdot 2^{2k}$, however, we observe a further error decay. Since $g_{\mathrm{p}} \in \bar{H}^s$ and $g_{\mathrm{p}} \in \bar{H}^s_{\mathrm{mix}}$ for all $s > 0$, we could expect super-algebraic convergence of the discretization error, see also (4.54) and (4.57). However, due to the early dominance of the sampling error already for $k > 3$, this cannot be observed in the plots in the top row of figures 6.8 and 6.9. For $k \geq 4$, the oversampling effect is already visible, i.e. the error increases again.

For fixed k and varying n, i.e. in the bottom row of the figures, we again observe the rate $\frac{n}{\log(n)}$ if k and n are large enough. The error decay for small n, where the noise level is not yet met by the overall error, is faster. In contrast to the non-periodic case, however, we cannot easily determine the decay rate here since it is different for each choice of k.

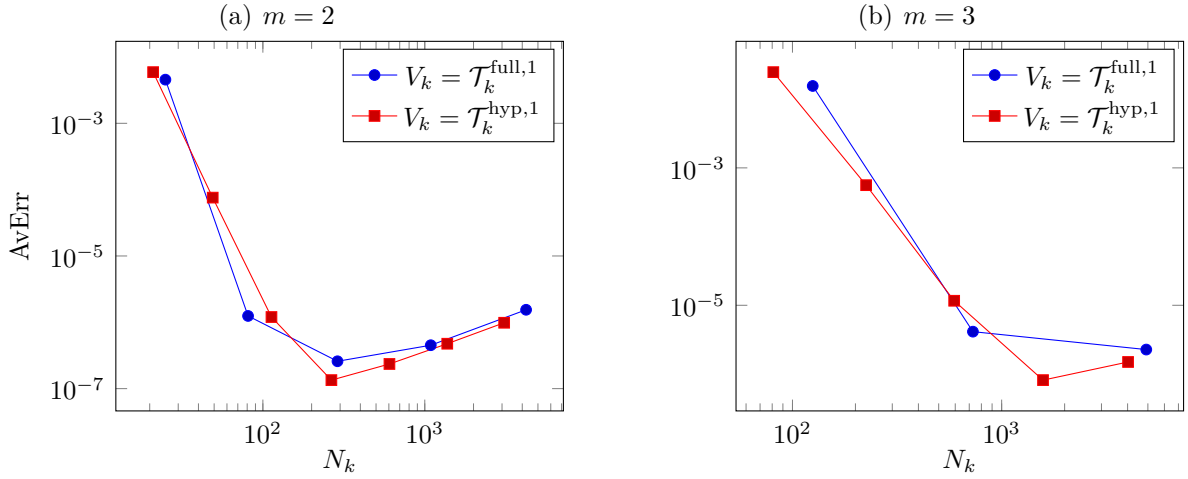

Fig. 6.10: The average overall error AvErr in dependence on the basis size N_k for full grids and hyperbolic crosses with $n = 2^{18}, \mu = 10^{-3}2^{-k}$. We plotted the two-variate case $m = 2$ (left) and the three-variate case $m = 3$ (right).

The overall error in dependence on N_k

Both the full grid and the hyperbolic cross discretizations lead to super-algebraic convergence of the discretization error. For the specific example g_{p}, we do not observe a significant difference between the discretization error behavior with respect to the degrees of freedom N_k for the full grid and the hyperbolic cross, see figure 6.10.

Stability

As we have shown in section 5.5, we obtain $K(N_k) = N_k$ for Fourier polynomials on full grids and hyperbolic crosses. Therefore, the oversampling factor c in $cN_k \leq \frac{n}{\log(n)}$ needs to be larger than

$$\frac{(1+\sigma)}{\lambda_{\min}(M) \cdot |\log(c_{\frac{1}{2}})|}$$

to apply theorem 5.11 and obtain a uniform stability bound. Since the Fourier polynomials are L_2-orthonormal, we have $\lambda_{\min}(M) = 1$. Choosing $\sigma = 1$, we obtain that $c \geq 18.48 \approx 2^{4.2}$ and, thus, $2^{4.2}N_k \leq \frac{n}{\log(n)}$ should suffice. This is essentially $4.2 + 2k \leq \log_2(n)$ for full grids and $4.2 + k \leq \log_2(n)$ for hyperbolic crosses.

Having a closer look at the computed average condition numbers for the system matrix $S = G + \mu C$, which we plotted in the figures 6.11 and 6.12, we observe that the full grid oversampling condition quite accurately describes the area for which $\kappa(S) \leq 10$. In the case of hyperbolic crosses, the condition $4.2 + k \leq \log_2(n)$ is fulfilled also by several pairs (k, n) for which $\kappa(S) > 10$. However, the corresponding condition numbers are still smaller than 10^3 for CC and LC, and they are smaller than 10^5 for QC. Thus, all

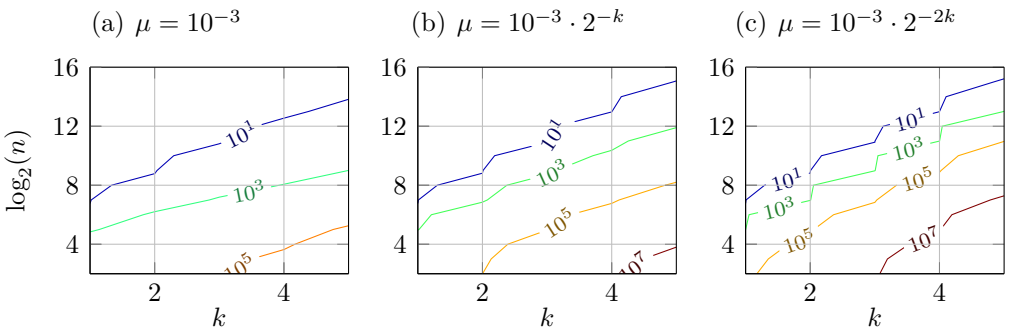

Fig. 6.11: Contour plot of the average of the condition numbers $\kappa(G + \mu C)$ over 10 independent draws of input data $\mathcal{Z}_n$ for full grids $V_k = \mathcal{T}_k^{\text{full},1}$. We depicted the contour lines for $\kappa = 10, 10^3, 10^5$ and 10^7.

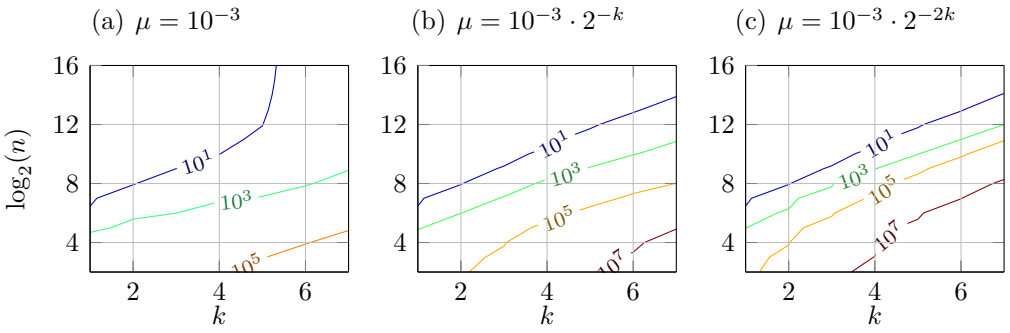

Fig. 6.12: Contour plot of the average of the condition numbers $\kappa(G + \mu C)$ over 10 independent draws of input data $\mathcal{Z}_n$ for hyperbolic crosses $V_k = \mathcal{T}_k^{\text{hyp},1}$. We depicted the contour lines for $\kappa = 10, 10^3, 10^5$ and 10^7.

systems are still treatable. As in the non-periodic case, we see that the transition area between well-conditioned and ill-conditioned systems shrinks from CC over LC to QC.

Also for the periodic case, we can conclude that the linear coupling $\mu = 10^{-3}2^{-k}$ is the best choice to achieve a small error and still lead to treatable systems for many choices of k and n.

Balanced convergence rate

As mentioned before, we could theoretically choose the smoothness parameter $s > 0$ arbitrarily large for this example. For the balancing between the discretization error and the sampling error, this implies that the larger the oversampling, the closer the convergence rate should get to n^{-1}, see also table 4.2 (b). However, because we only have finite computing resources at hand, there is a limit to the amount of data n which can be treated. Note that, in subsection 6.1.1, we looked at the optimal coupling between

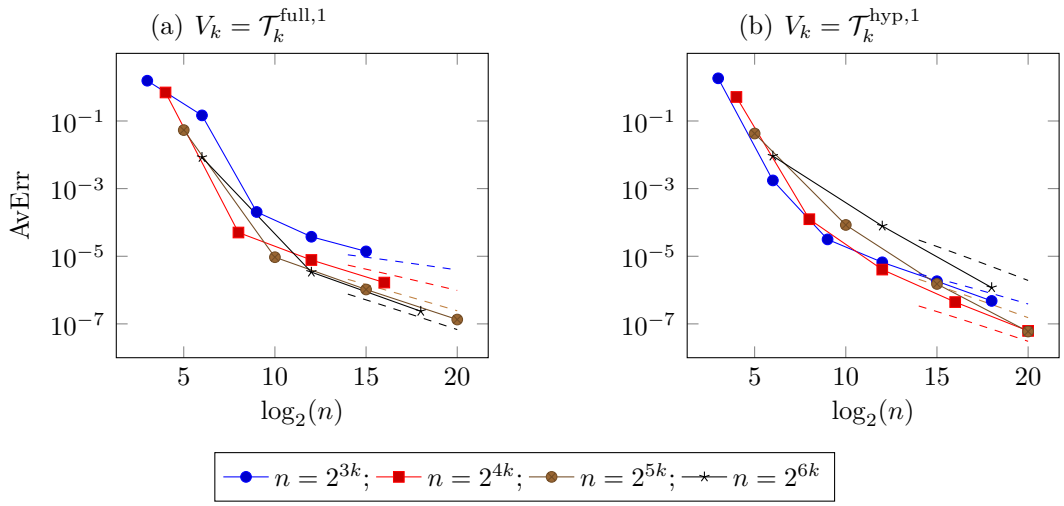

Fig. 6.13: AvErr for full grids (left) and hyperbolic crosses (right) with $\mu = 10^{-3}2^{-k}$. We balanced the discretization error and the sampling error by using the relation $n = 2^{lk}$ with $l \in \{3, 4, 5, 6\}$. The upper bounds which we derived in chapter 4, i.e. $n^{-\frac{l-2}{l}} \log(n)$ for full grids and $n^{-\frac{l-1}{l}} \log(n)^2$ for hyperbolic crosses, are plotted as dashed lines with color according to the corresponding coupling. These rates are obtained by setting $n = 2^{lk} = 2^{(2s+m)k}$ for full grids and $n = 2^{lk} = 2^{(2s+1)k}$ for hyperbolic crosses and taking the corresponding result from table 4.2 (b).

n and k in terms of the overall convergence rate with respect to n. Here, this is not possible since we would have to choose an infinite oversampling. Hence, we compute the results for different choices of the coupling between n and k instead and compare them.

In figure 6.13, we observe the error behavior for couplings of type $n = 2^{lk}$ with $l = 3, 4, 5, 6$. For full grids, the observed convergence behavior approximately matches our upper bounds for $l = 4, 5, 6$. In the hyperbolic cross case, our theoretical bounds seem to be too pessimistic, especially for $l = 5, 6$. However, it is not clear if the bounds are indeed not sharp or if we are observing only preasymptotic behavior in our plots.

6.2 Convergence analysis for unregularized regression of noiseless data

We now consider noiseless function regression with the same test functions as in the previous section. Thus, we have $\rho(\mathbf{x}|\mathbf{t}) = \delta_{g_*(\mathbf{t})}(\mathbf{x})$, where $\delta_{g_*(\mathbf{t})}$ denotes the Dirac distribution centered in $g_*(\mathbf{t})$ and $* \in \{\text{p,np}\}$. We deal with the unregularized case $\mu = 0$ in our computations, which means that the system matrix $S = G + \mu C = G = nB \circ B^*$ cannot have full rank if $n < N_k$. Therefore, we only consider parameter tuples (k, n)

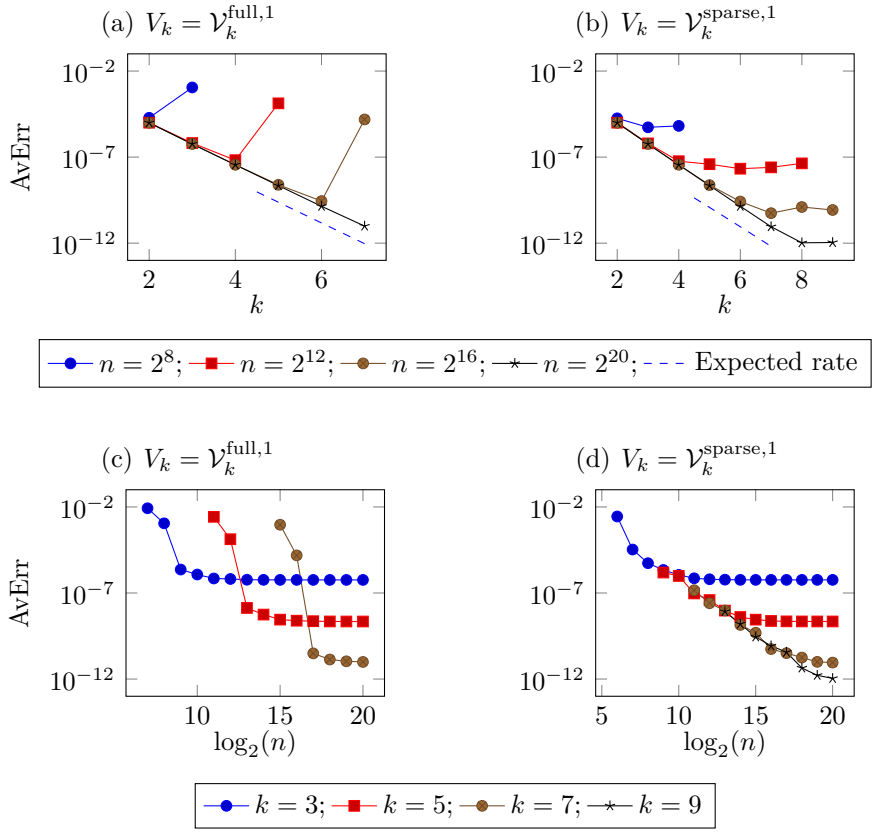

Fig. 6.14: The average AvErr of the overall error $\text{Err} = \mathcal{E}\left(f_{\mathcal{Z}_n,V_k,\mu}\right) - \mathcal{E}\left(g_{\text{np}}\right)$ over 10 independent draws of input data $\mathcal{Z}_n$ for the function g_{np} is plotted versus the level k for different choices of n (top) and versus the number of data n for different choices of k (bottom). Both full grid (left) and sparse grid (right) results are depicted. The expected rate in the top row is 2^{-4k} in the full grid case and $2^{-4k} \cdot k$ in the sparse grid case.

with $N_k \leq n$. We are now in a setting where we can employ our results on noiseless function regression from section 5.5.

6.2.1 Non-periodic regression on full grids and sparse grids

Our setup is exactly the same as in the previous section, apart from the fact that we now deal with noiseless samples of g_{np} and $\mu = 0$.

The overall error in dependence on k and n

We plotted the error for different choices of k and n in figure 6.14. In the top row, we observe that the overall error in dependence on k decreases with the expected rate

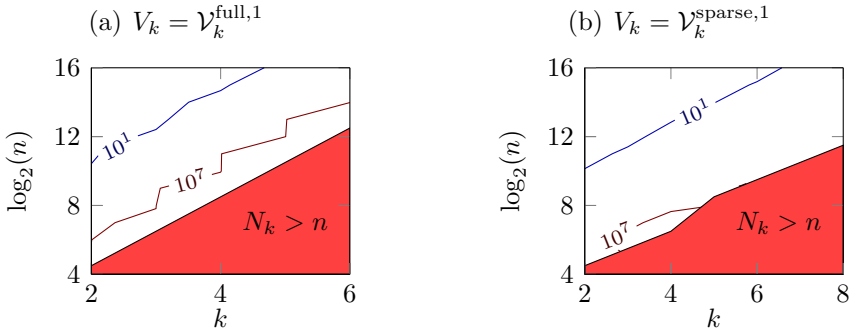

Fig. 6.15: Contour plot of the average of the condition numbers $\kappa(G)$ over 10 independent draws of input data $\mathcal{Z}_n$ for full grids (left) and sparse grids (right). We depicted the contour lines for $\kappa = 10$ and $\kappa = 10^7$.

of 2^{-4k} for full grids and $2^{-4k} \cdot k$ for sparse grids. If k gets too large in comparison to n, the error either stagnates (sparse grid) or even increases (full grid). The latter is due to the instability in the case $N_k \approx n$ for full grids, which we will observe in the follow-up subsection. A related phenomenon can be discovered when inspecting the error for varying n and fixed k (bottom row of figure 6.14). For full grids, the error is very large when $N_k \approx n$, but it rapidly drops to a plateau when n increases, i.e. when the discretization error begins to dominate. For sparse grids, the error in the case $N_k \approx n$, i.e. at the first dot of each colored line, is significantly smaller for $k \geq 5$.

Stability

We depicted the condition number $\kappa(G)$ of the system matrix G in dependence on k and n in figure 6.15. Here, we plotted the contour lines for $\kappa(G) = 10$ and $\kappa(G) = 10^7$. Recall that these contour lines characterize the transition between well-conditioned, treatable and ill-conditioned systems according to our earlier definition in section 6.1. Furthermore, we filled the region in which $N_k > n$ holds with a red color. In the full grid case, we observe that the problem becomes ill-conditioned, i.e. $\kappa(G) \geq 10^7$, if N_k is close to n. This explains why the error in figure 6.14 (a) increases again when k becomes too large in comparison to n. In the sparse grid case, however, all systems are well-conditioned or at least treatable, i.e. $\kappa(G) < 10^7$, except for the ones with parameter pairs $k = 3$, $n = 2^6$ and $k = 4$, $n = 2^7$.

Since the contour line for $\kappa = 10$ is approximately the same as in the regularized example from subsection 6.1.1, the application of theorem 5.11 gives again a good estimate for adequate parameter choices which lead to well-conditioned systems.

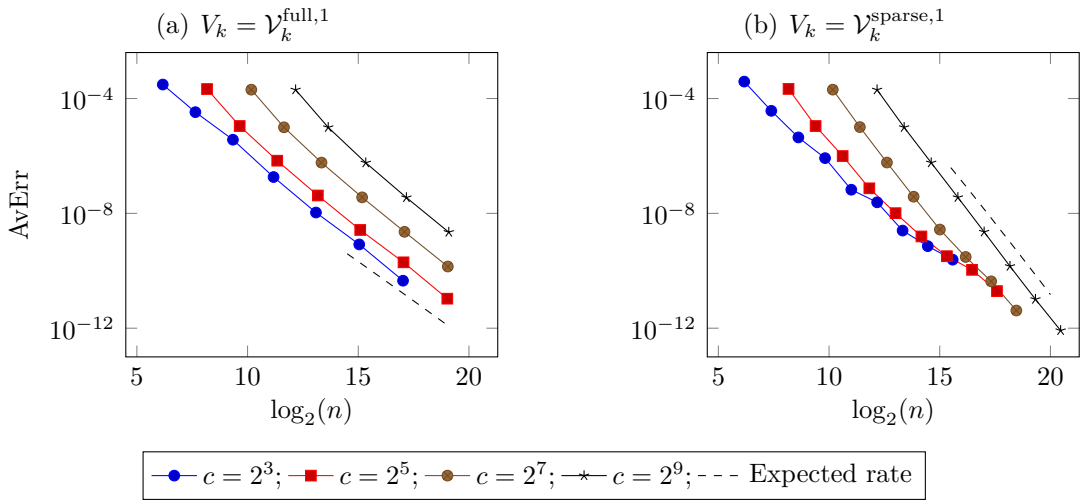

Fig. 6.16: The average overall error AvErr for full grids (left) and sparse grids (right) after balancing the summands in the overall error bound. We used the coupling $N_k = n \cdot c$, for an offset $c > 0$. If the offset is large enough, we expect the rates $n^{-2}\log(n)^2$ for full grids and $n^{-4}\log(n)^9$ for sparse grids, see table 5.2 (a) for $m = s = 2$.

Balanced convergence rate

To simulate the balanced case, we choose[2] the sample size to be $n = N_k \cdot c$ for an offset factor $c > 0$, which is directly related to the oversampling constant σ from (5.11). Clearly, if c is chosen too small, we cannot expect to get the optimal rates derived in section 5.5. However, if c is large enough to ensure $\sigma \geq \frac{2s}{m} = 2$ for full grids and $\sigma \geq 2s = 4$ for sparse grids, we expect to observe convergence rates of at least

$$n^{-\frac{2s}{m}} \log(n)^{-\frac{2s}{m}} = n^{-2}\log(n)^2$$

for full grids and

$$n^{-2s}\log(n)^{(2s+1)m-1} = n^{-4}\log(n)^9$$

for sparse grids, see table 5.2 (a). We plotted our results for different choices of the offset c in figure 6.16. For full grids, the expected rate is met already for $c = 2^3$. For sparse grids, the expected rate can be observed for $c = 2^9$. Note that, c naturally influences the constant in the rates. To achieve a fixed error, it might, therefore, be better to choose a smaller c, even though the asymptotic convergence rate is worse.

[2]Note that $\frac{n}{\log(n)} = N_k \cdot c$ would be more appropriate to reflect the balancing from table 5.2 in the asymptotic case. However, since we only deal with n up to 20 here, we can ignore the logarithm.

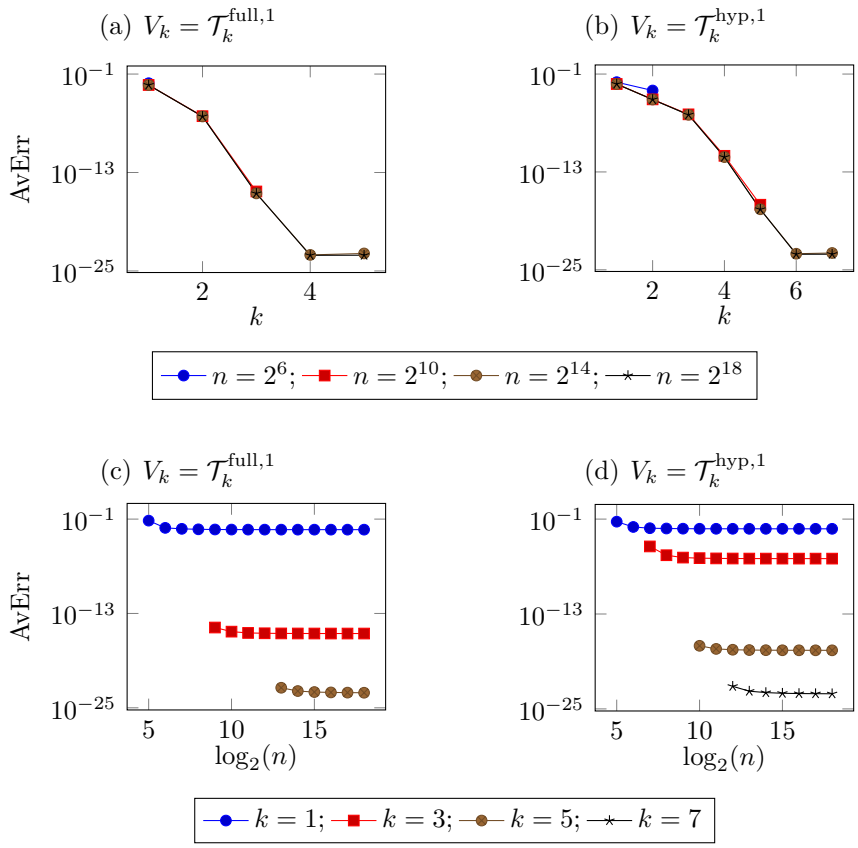

Fig. 6.17: The average AvErr of the overall error $\text{Err} = \mathcal{E}\left(f_{\mathcal{Z}_n,V_k,\mu}\right) - \mathcal{E}\left(g_{\text{p}}\right)$ over 10 independent draws of input data $\mathcal{Z}_n$ for the function g_{np} is plotted versus the level k for different choices of n (top) and versus the number of data n for different choices of k (bottom). Both full grid (left) and hyperbolic cross (right) results are plotted.

6.2.2 Periodic regression on full grids and hyperbolic crosses

In this subsection, we consider unregularized regression ($\mu = 0$) of the periodic function g_{p} in the noiseless setting.

The overall error in dependence on k and n

In figure 6.17, we plotted the error in dependence on k and n. As we observe, the qualitative behavior is opposite to the one we witnessed for noisy function regression in figures 6.8 and 6.9. There, the discretization error quickly reached its minimum and the overall error was governed by the rate of convergence of the sampling error. In figure 6.17, however, the error with respect to n is almost constant and convergence can only be seen when increasing k. Note that 10^{-24} is the machine precision and we cannot

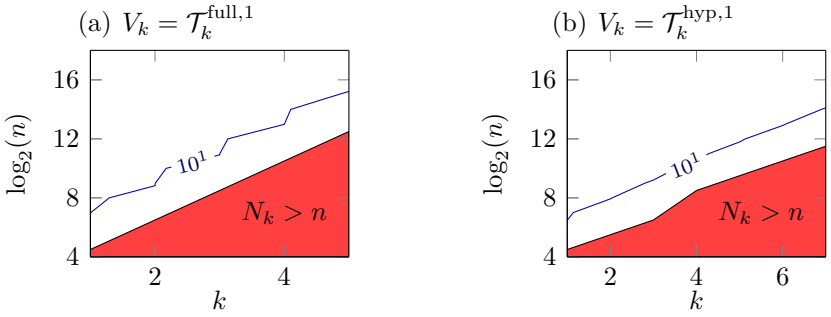

Fig. 6.18: Contour plot of the average of the condition numbers $\kappa(G)$ over 10 independent draws of input data $\mathcal{Z}_n$ for full grids (left) and hyperbolic crosses (right). We depicted the contour line for $\kappa = 10$. All calculated condition numbers are smaller than 10^4.

expect to observe any further decay of the error. Having a closer look at the error bounds (5.33) and (5.36), which we derived for noiseless regression on full grids and hyperbolic crosses, our experimental results have to be interpreted as follows: Although we theoretically have super-algebraic decay of the k-dependent term 2^{-2sk} since $s > 0$ can be chosen arbitrarily large, the constants involved there are such that the corresponding summand is still larger than the data dependent term governed by $n^{-\sigma}$ already for moderate oversampling parameters $\sigma > 0$, i.e. already for n close to N_k.

Stability

A contour plot of the average condition numbers $\kappa(G)$ of the system matrix G for unregularized regression in the periodic case can be found in figure 6.18. We depicted the contour line $\kappa(G) = 10$ and marked the area in which $N_k > n$ holds. Note that all of the computed condition numbers were smaller than 10^4 and the contour line for $\kappa(G) = 10^7$, which we plotted in figure 6.15, cannot be seen here. Therefore, all systems are well-conditioned or at least treatable.

We observe that the contour line for $\kappa = 10$ approximately matches the corresponding line in the LC case of the noisy regression problem, see also figure 6.12. Therefore, everything we discussed about the application of theorem 5.11 in subsection 6.1.2, is also valid here.

Balanced convergence rate

Similarly as in the non-periodic case, we take $n = N_k \cdot c$ for different choices of $c > 0$. Since $s > 0$ can be chosen arbitrarily large, we could expect that the convergence becomes faster with increasing c in the balanced case, see table 5.2. However, as we already saw in figure 6.17, the constants involved in the 2^{-2sk} term of the error bound and the finite

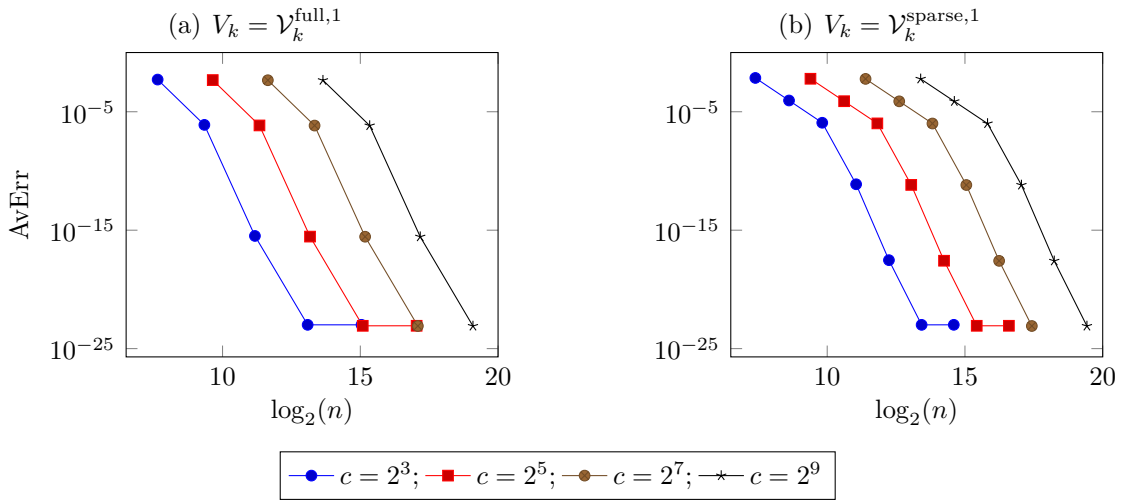

Fig. 6.19: The average overall error AvErr for full grids (left) and hyperbolic crosses (right) after balancing the summands in the overall error bound. We used the coupling $N_k = n \cdot c$, for an offset $c > 0$.

machine precision prevent us from observing arbitrarily high convergence rates. We plotted the experimental results for the balanced case in figure 6.19. There, we see that the convergence rate is not influenced by the choice of c. However, as we mentioned, this is most probably due to the fact that we cannot compute the error with arbitrary high precision and that the computational resources limit us to the case $k < 10$, $n < 2^{20}$. Nevertheless, we observe convergence rates faster than n^{-15} already in figure 6.19.

6.3 Adaptivity

In contrast to the previous sections, we now consider a non-smooth example function. To this end, let the vector-valued function $f_{\text{jump}} : (0,1)^5 \to \mathbb{R}^3$ be defined by

$$f_{\text{jump}}(t_1, \ldots, t_5) := \left(\chi_{[0.4,1)}(t_1),\ \chi_{[0.4,1)}(t_2),\ \sum_{i=1}^{5} t_i \right)^T,$$

where

$$\chi_{[0.4,1)}(t) := \begin{cases} 1 & \text{if } t \in [0.4, 1) \\ 0 & \text{else} \end{cases}$$

denotes the characteristic function on the interval $[0.4, 1)$. Since the first two components of f_{jump} contain a jump, the smoothness of this function is limited. Indeed, it can be shown that we only have $f_{\text{jump}} \in H_{\text{mix}}^{\frac{1}{2}-\delta}$ for any $\delta > 0$. Therefore, when drawing samples from $g = f_{\text{jump}}$ and running a sparse grid regression algorithm, we can expect our

convergence results to hold only with smoothness parameter $s < \frac{1}{2}$. Thus, the squared best approximation error $\inf_{f \in \mathcal{V}_k^{\text{sparse},1}} \|f_{\text{jump}} - f\|^2_{L_2((0,1)^5;\mathbb{R}^3)}$ can be expected to decay with a rate of only $2^{-2sk} \cdot k^{m-1} > 2^{-k} k^4$, see also (3.42). This, of course, directly carries over to the decay of the discretization error in (4.52) in the noisy case or the k-dependent summand in (5.29) in the noiseless case, respectively.

The slow rate of convergence of the best approximation error is a major drawback when applying regression algorithms. In the noiseless case for instance, the convergence with respect to n is very fast. Therefore, the k-dependent error term is usually dominant in the overall error bound, see also section 6.2. In order to achieve a certain overall error, the basis size N_k needs to be significantly larger in the case of a non-smooth g than in the case of a smooth solution.

Adaptivity can be employed to remedy this problem to some extent. In the following, we briefly introduce a dimension-adaptive variant of the sparse grid regression algorithm based on [30] and investigate its performance for our example. The idea behind the algorithm is to refine the resolution of the discretization in the most important coordinate directions and coarsen it in the other ones. For our example, this means that e.g. the first component function $\chi_{[0.4,1)}(t_1)$ needs to be resolved quite accurately in t_1 direction. However, since it does not depend on the other coordinates at all, the resolution in these directions can be coarsened.

6.3.1 A dimension-adaptive sparse grid regression algorithm

In order to describe the dimension-adaptive sparse grid algorithm, we have to make a few alterations to our previous definitions for sparse grids from section 3.5. For detailed explanations, we refer to [11, 30]. Let the altered index set $\tilde{\mathbf{I}}_{\mathbf{l}}$ for an $\mathbf{l} \in (\mathbb{N} \cup \{-1\})^m$ be defined by

$$\tilde{\mathbf{I}}_{\mathbf{l}} := \left\{ \mathbf{i} \in \mathbb{N}^m \;\middle|\; \begin{array}{ll} i_j = 0, & \text{if } l_j = -1, \\ i_j = 1, & \text{if } l_j = 0, \\ 1 \leq i_j \leq 2^{l_j} - 1,\ i_j \text{ odd} & \text{if } l_j > 0, \end{array} \quad \text{for } 1 \leq j \leq m \right\}.$$

Furthermore, let the univariate basis function $\gamma_{-1,0}(t) := 1$ be constant on $[0, 1]$ and let us redefine $\gamma_{0,1}(t) := 2t - 1$. With the definition

$$\tilde{W}_{\mathbf{l}} := \operatorname{span}\{\gamma_{\mathbf{l},\mathbf{i}} \mid \mathbf{i} \in \tilde{\mathbf{I}}_{\mathbf{l}}\},$$

compare (3.31), we directly obtain $W_{\mathbf{l}} = \tilde{W}_{\mathbf{l}}$ for every $\mathbf{l} \in (\mathbb{N} \setminus \{0\})^m$. By using

$$\tilde{\zeta}_m(\mathbf{l}) := \begin{cases} 0 & \text{if } l_j \leq 0 \text{ for all } 1 \leq j \leq m, \\ \sum_{j=1}^m \max(l_j, 0) - m + |\{j \mid \mathbf{l}_j \leq 0\}| + 1 & \text{else} \end{cases}$$

instead of $\zeta_m(\mathbf{l})$, we obtain the equality

$$\mathcal{V}_k^{\text{sparse},d} = \bigoplus_{\substack{\mathbf{l}\in(\mathbb{N}\cup\{-1\})^m \\ \tilde{\zeta}_m(\mathbf{l})\leq k}} \tilde{W}_{\mathbf{l}}$$

for the sparse grid spaces with $k > 0$. However, the specific choice of $\tilde{W}_{\mathbf{l}}$ now establishes a direct link of the decomposition of the sparse grid space discretization to the so-called *analysis-of-variance* (*ANOVA*) decomposition. For a detailed introduction of the ANOVA decomposition and a thorough explanation of the mentioned link, we refer the interested reader to [11, 30].

The error indicator

The main ingredient for the dimension-adaptive sparse grid regression algorithm is the error indicator. Its size determines if the grid is refined in a certain direction. Let $\Xi_i \subset (\mathbb{N}\cup\{-1\})^m$ be arbitrary *lower* multilevel index sets for the directions $i = 1,\ldots,d$. As we already mentioned in subsection 5.5.6, the term "lower" means that $\mathbf{l} \in \Xi_i$ implies $\mathbf{k} \in \Xi_i$ for all $\mathbf{k} \leq \mathbf{l}$. Let, furthermore, $f : (0,1)^m \to \mathbb{R}^d$ be a function with components $f_i, i = 1,\ldots,d$, given in the form

$$f_i = \sum_{\mathbf{l}\in\Xi_i} \sum_{\mathbf{j}\in\tilde{\mathbf{I}}_{\mathbf{l}}} \alpha_{\mathbf{l},\mathbf{j},i}\, \gamma_{\mathbf{l},\mathbf{j}}.$$

The error indicator $\epsilon_{\mathbf{l},i}$ for the $\mathbf{l}$-th multilevel index of the i-th component function of f is defined by

$$\epsilon_{\mathbf{l},i}(f) := \max_{\mathbf{j}\in\tilde{\mathbf{I}}_{\mathbf{l}}} \|\alpha_{\mathbf{l},\mathbf{j},i}\,\gamma_{\mathbf{l},\mathbf{j}}\|_{L_2([0,1]^m)} \simeq \max_{\mathbf{j}\in\tilde{\mathbf{I}}_{\mathbf{l}}} \alpha_{\mathbf{l},\mathbf{j},i}.$$

We refer to [38] for details on this specific error indicator. To see that this choice is meaningful, let us consider the regression error $\mathcal{E}(f) - \mathcal{E}(g)$. Let the components g_i of g with $i = 1,\ldots,d$ be given by

$$g_i := \sum_{\mathbf{l}\in(\mathbb{N}\cup-1)^m} \sum_{\mathbf{j}\in\tilde{\mathbf{I}}_{\mathbf{l}}} \beta_{\mathbf{l},\mathbf{j},i}\, \gamma_{\mathbf{l},\mathbf{j}}.$$

If f is a good approximation to g, e.g. if f is the solution to the regression problem, then $\alpha_{\mathbf{l},\mathbf{j},i} \approx \beta_{\mathbf{l},\mathbf{j},i}$ for all $i = 1,\ldots,d$ and $\mathbf{l} \in \Xi_i, \mathbf{j} \in \tilde{\mathbf{I}}_{\mathbf{l}}$. Therefore, we obtain

$$\mathcal{E}(f) - \mathcal{E}(g) = \|f - g\|^2_{L_2((0,1)^m;\mathbb{R}^d)} \simeq \sum_{i=1}^{d} \sum_{\mathbf{l}\notin\Xi_i} \sum_{\mathbf{j}\in\tilde{\mathbf{I}}_{\mathbf{l}}} \beta^2_{\mathbf{l},\mathbf{j},i}$$

according to (3.38) with $s = 0$. Thus, the overall error is reduced the most, when adding those $\mathbf{l}$ to Ξ_i for which $\sum_{\mathbf{j}\in\tilde{\mathbf{I}}_{\mathbf{l}}} \beta^2_{\mathbf{l},\mathbf{j},i}$ is the largest. Since the coefficients of $g \in H^s_{\text{mix}}$ have to decay quickly with increasing multilevel index, see again (3.38), it is reasonable to

assume that $\beta^2_{\mathbf{k},\mathbf{j},i} < \beta^2_{\mathbf{l},\mathbf{j},i}$ for $\mathbf{k} > \mathbf{l}$. Therefore, we do not need to refine further into directions $\mathbf{l}$ for which $\epsilon_{\mathbf{l},i}(f)$ is already small. Note, however, that this is of course only a heuristic approach and there exist counter-examples where this idea does not work, see e.g. [38] for details.

The dimension-adaptive algorithm

The dimension-adaptive procedure to compute the solution $f_{\mathcal{Z}_n,A,\mu}$ of the regression problem on an adapted grid space A can be found in algorithm 1. The input parameters are the threshold $\varepsilon > 0$, the maximum refinement level $k_{\text{end}} > 0$ and the initial lower multilevel index sets $\Xi_i \subset (\mathbb{N} \cup \{-1\})^m$ for $i = 1, \ldots, d$. Prior to the adaption, we perform a compression step in order to detect directions which are irrelevant for our further computations. In the refinement step, we consider the directions $\mathbf{l} + \mathbf{e}_j$ for all $j \in \{1, \ldots, m\}$ for which $l_j \neq -1$. This means that we exclude the refinement of directions in which f is constant since they are not important for the representation of f. For more mathematical details on this step, we refer to [11, 30]. Note that we run over all $\mathbf{k} \leq \mathbf{l} + \mathbf{e}_j$ to ensure that the index sets of the underlying grid are lower also after the refinement. This is needed to ensure that a grid traversal algorithm works correctly, see [29, 30] for technical details.

6.3.2 Regular sparse grids vs. dimension-adaptive sparse grids

We now compare the performance of our standard algorithm for (regular) sparse grids to the dimension-adaptive algorithm. To this end, we take $g = f_{\text{jump}}$ and randomly sample $n = 2^{15}$ data points. Since we investigate the noiseless case, i.e. $\rho(\mathbf{x}|\mathbf{t}) = \delta_{g(\mathbf{t})}(\mathbf{x})$, we use $\mu = 0$. For the numerical error measurement, we interpolated g on a standard sparse grid of level 6. In figure 6.20, we plotted the overall error for regular sparse grids and dimension-adaptive sparse grids with respect to the basis size. We started with a grid over the index sets $\Xi_i = \{-1, 0\}^m$ in each direction $i = 1, \ldots, d$ and varied the k_{end} parameter.

First of all, we notice that the adaptive algorithm clearly outperforms the regular sparse grid method for both choices of ε. In the case $\varepsilon = 10^{-2}$, we observed during the experiment that the algorithm detects the structure of the function perfectly, i.e. the first component is coarsened up to two remaining grid points and, subsequently, refined only in t_1 direction. The same holds true for the second component and the t_2 direction. The algorithm also detects that the sum $\sum_{i=1}^5 t_i$ in the third component can be described by the basis functions with the multilevel indices $\mathbf{l} = (-1, -1, -1, -1, -1)^T + \mathbf{e}_j$ for each $j = 1, \ldots, 5$. If we choose ε too small, the grid is also refined in several irrelevant directions as we observe for $\varepsilon = 10^{-3}$. However, also in this case the algorithm still outperforms the regular sparse grid method. It is noteworthy that the runtime of the regular sparse grid computation for level $k = 3$ was more than 10 times larger than the runtime for each of the adaptive calculations.

Algorithm 1 Computation of $f = f_{\mathcal{Z}_n,A,\mu}$ in the dimension-adaptive space A

Require: $\varepsilon > 0$, $k_{\text{end}} > 0$, $\Xi_i \subset (\mathbb{N} \cup \{-1\})^m$ for $i = 1, \ldots, d$
Ensure: f: the solution to the regression problem on the refined grid

```
A := (A_1, ..., A_d)^T with A_i ← ⊕_{l∈Ξ_i} W̃_l for i = 1, ..., d   {Initialization of A}
f ← f_{Z_n,A,μ}   {Initialization of f}
for i = 1, ..., d do
  for W̃_l ⊂ A_i do
    if ε_{k,i}(f) ≤ ε‖f_i‖_{L_2([0,1]^m)} for all k ∈ Ξ_i with l ≤ k then
      A_i ← A_i \ {W̃_l}   {Initial compression}
    end if
  end for
end for
while |l|_{ℓ_∞} < k_end for all l with W̃_l ⊂ A_i for an i ∈ {1, ..., d} do
  f ← f_{Z_n,A,μ}   {Calculate solution on actual grid}
  for i = 1, ..., d do
    for W̃_l ⊂ A_i do
      if ε_{l,i}(f) > ε‖f_i‖_{L_2([0,1]^m)} then
        for j ∈ {1, ..., m} with l_j ≠ −1 do
          for k ≤ l + e_j do
            A_i ← A_i ∪ W̃_k   {Refinement step}
          end for
        end for
      end if
    end for
  end for
end while
f ← f_{Z_n,A,μ}   {Calculate the solution on the refined space A}
return f
```

In summary, the dimension-adaptive variant has two main advantages over its regular counterpart: First, we obtain significantly improved error convergence rates with respect to N_k for anisotropic problems. Second, we can employ a much higher grid resolution in important directions of the function before we encounter the critical $N_k \approx n$ barrier. Besides the runtime, this also affects the stability of the corresponding equations.

We cannot expect the adaptive algorithm to perform as nicely as for our toy example when dealing with more complicated problems. However, there exist ideas on how to extend the class of problems which can be successfully treated by the dimension-adaptive sparse grid algorithm. By applying an initial rotation to the data for instance, functions with jumps along diagonal directions can also be handled efficiently, see [65] for first results in this direction. Besides the sparse grid case, the general ideas of this section

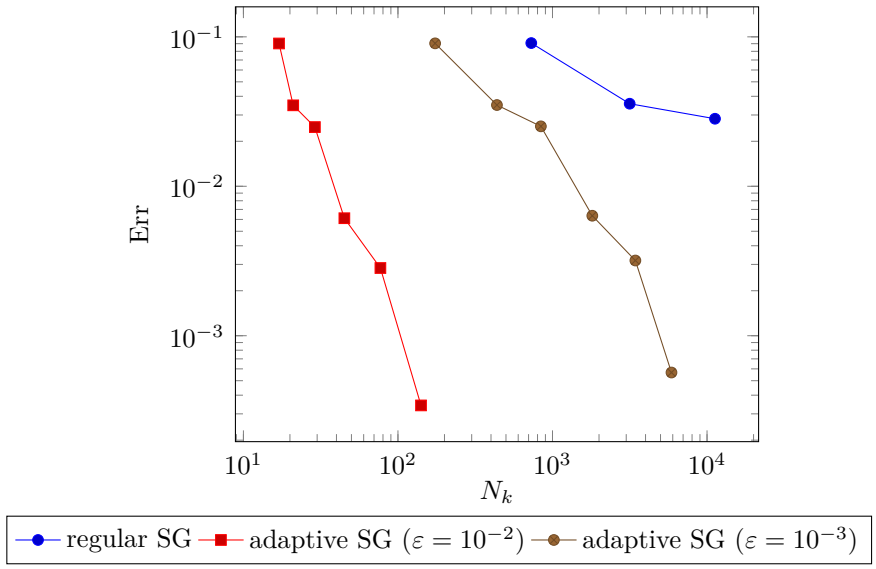

Fig. 6.20: The overall error $\mathrm{Err} := \mathcal{E}(f_{\mathcal{Z}_n,V_k,0}) - \mathcal{E}(f_{\mathrm{jump}})$ in dependence on the basis size N_k. Here, $V_k = \mathcal{V}_k^{\mathrm{sparse},5}$ for regular sparse grids and $V_k = A$ with $k = k_{\mathrm{end}}$ for dimension-adaptive sparse grids. The plotted levels for the regular SG are $k = 1, 2, 3$. For the dimension-adaptive SG we start with $\Xi_i = \{-1, 0\}^m$ for all $i = 1, \ldots, 5$ and plot the results for $k_{\mathrm{end}} = 1, 2, 3, 4, 5, 6$ with $\varepsilon = 10^{-2}$ and $\varepsilon = 10^{-3}$.

can also be transferred to hyperbolic cross regression.

6.4 Real world examples

In this section, we show how the dimension-adaptive algorithm can be used to tackle problems which stem from real-world applications. To this end, we consider a data set of electroencephalogram (EEG) measurements in one experiments and the results of professional soccer games over the last 8 years in a second experiment. In both cases, the number of input data points is quite large and exceeds 10^4. It is unclear, however, if the data at hand fulfills the assumptions posed in our earlier chapters, e.g. independence of measurements, or if the true solution resides in a Sobolev space of mixed smoothness. Since the data points $\mathbf{t}_1, \ldots, \mathbf{t}_n$ do not necessarily reside in $(0, 1)^m$ in the experiments, we rescale the domain linearly in each coordinate direction in our code.

6.4.1 Eye state prognosis from EEG measurements

The goal in this experiment is to correctly predict the eye state (open or closed) of a test subject from given EEG data.

The data set

The "EEG eye state data set" is taken from the UCI machine learning repository, see [57]. It provides 117 seconds of EEG measurements from one single proband. More specifically, it consists of 14-dimensional points $\mathbf{t}_i$ for $i = 1, \ldots, 14980$, where each data point corresponds to a fixed time step and represents the values measured by the $m = 14$ electrodes of the electroencephalograph. Furthermore, we are given $\mathbf{x}_i \in \{-1, 1\}$, which indicate if the eyes of the test subject were open (-1) or closed (1) at time step i. Here, approximately 55% of the data have the label -1 and 45% carry the label 1. Since there is obvious measurement noise in 4 of the data points, we remove them to obtain a data set with 14976 points $(\mathbf{t}_i, \mathbf{x}_i)$.

Learning the data

To tackle this problem with our dimension-adaptive sparse grid regression algorithm, we perform a 10-fold crossvalidation, i.e. we split the set $\{1, \ldots, 14976\}$ into 10 parts of (approximately) equal size. Then, we use the union of 9 of these parts as training indices I_{train} to learn a model $f_{\mathcal{Z}_n,A,\mu}$ for which all data points $\mathcal{Z}_n = (\mathbf{t}_i, \mathbf{x}_i)_{i \in I_{\text{train}}}$ are taken as input data. We evaluate the classification rate

$$\text{ClassRate}(f_{\mathcal{Z}_n,A,\mu}) := \frac{1}{|I_{\text{test}}|} \sum_{i \in I_{\text{test}}} \text{Class}(f_{\mathcal{Z}_n,A,\mu}(\mathbf{t}_i), \mathbf{x}_i)$$

with

$$\text{Class}(f_{\mathcal{Z}_n,A,\mu}(\mathbf{t}_i), \mathbf{x}_i) := \begin{cases} 1 & \text{if } f_{\mathcal{Z}_n,A,\mu}(\mathbf{t}_i) \cdot \mathbf{x}_i > 0 \\ 0 & \text{else} \end{cases}$$

on the test data corresponding to $I_{\text{test}} = \{1, \ldots, 14976\} \setminus I_{\text{train}}$. Subsequently, we take another one of the 10 parts of the index set as I_{test} and learn a model from the remaining 9 parts. This process is repeated until each of the 10 parts has been chosen as I_{test} once. The arithmetic average of the classification rate over all 10 runs serves as a measurement for the performance of our algorithm and allows to directly compare our results to the ones in [71].

We could think about solving the whole 14-dimensional problems with a rather large refinement threshold ε to limit the size of the resulting grid. However, besides the high computational costs, this does not allow to refine the most important directions sufficiently well. Therefore, we choose a two-step approach:

- First, we tackle the 14-dimensional problem with the dimension-adaptive algorithm with $\varepsilon = 0.1$, $k_{\text{end}} = 3$, $\mu = 10^{-3}$ and an initial multilevel index set which consists of all 14-dimensional vectors from $\{-1, 0\}^{14}$ with at most three zero-entries. In this way, we only consider at most three-way interactions of the input variables.

- Afterwards, we take the three coordinate directions in which the adaptive algorithm spent the most amount of grid points and use them to start the adaptive algorithm with $\varepsilon = 0.01$, $k_{\text{end}} = 5$, $\mu = 10^{-3}$ and initial multilevel index set $\{-1, 0\}^3$ for the corresponding three-dimensional problem.

In this way, the adaptive algorithm first determines the most important directions and we then take only these coordinates to proceed with a fine-grained approach. In order to save computational time, we perform the initial 14-dimensional computation only for the first choice of I_{train} and keep the three most important dimensions fixed for our computations of the other folds. Note that the specific choice of the parameters of the dimension-adaptive sparse grid algorithm is rather ad hoc and has not been optimized in any way.

Evaluation

The three directions in which our dimension-adaptive method spends the most grid points correspond to three electrodes which were located at the left hemisphere of the brain. Here, one was placed to measure stimulus of the the occipital lobe, the second one measured the parietal lobe activity and the third one was placed at the posterior of the frontal lobe, close to the temporal lobe. Therefore, the algorithm decided not to consider the activity of the right hemisphere at all. Furthermore, it decided that combining the information from several lobes is more promising than relying on multiple measurements from a single lobe.

Our results can be found in table 6.1. The average classification rate is approximately 71.6%, which is of course too small for an actual clinical application, but it shows that there is a connection between the eye state and the EEG measurements. Since this 10-fold crossvalidation experiment has served as a benchmark for many machine learning algorithms, we can directly check how our algorithm ranks in comparison to others. As we see from the survey in [71], we outperform the majority of the classification algorithms tested there. Surprisingly, popular machine learning methods such as naive Bayes estimators, support vector machines and multilayer artificial neural networks seem to perform quite bad for this data set and rank significantly worse than our approach. More specifically, they achieve classification rates ranging from approximately 50% up to 68%. Note that the authors of [71] did not optimize the parameters for the corresponding methods - but neither did we. Our choices for ε, k_{end} and μ are quite generic.[3] Note, furthermore, that most of the methods that achieved a higher average classification rate than 71.6% in [71] are either based on a decision tree or on an instance-by-instance comparison. Thus, the underlying idea for these methods is quite different from our approach.

[3] We also tested different values here, but the results for our adaptive sparse grid method did not vary much.

test fold	classification rate
1	0.6947
2	0.7303
3	0.7128
4	0.7150
5	0.7256
6	0.7301
7	0.7096
8	0.7114
9	0.7210
10	0.7056
Average	0.7156

Table 6.1: The classification rates $\mathrm{ClassRate}(f_{\mathcal{Z}_n,A,\mu})$ for each test set in a 10-fold cross-validation of the EEG eye state data set. The calculations were performed on the three-dimensional data set which resulted from the initial analysis of the 14-dimensional problem. The parameters were chosen as $\varepsilon = 0.01$, $k_{\mathrm{end}} = 5$, $\mu = 10^{-3}$ and we started on the grid which contains all multilevel indices $\mathbf{l} \in \{-1, 0\}^3$.

6.4.2 Prediction of soccer matches

In this experiment, we investigate if it is possible to predict the tendential outcome of soccer matches, i.e. home team win, away team win or draw, based on player statistics from the "EA Sports FIFA" computer game series.

The data set

The corresponding "European Soccer Database", which we use here, can be found at `http://www.kaggle.com`. From this data set, we extracted the data for all matches from seasons 2008/2009 to 2015/2016 of the following leagues: Jupiler League (Belgium), Premier League (England), Ligue 1 (France), 1. Bundesliga (Germany), Serie A (Italy), Eredivisie (Netherlands), Liga ZON Sagres (Portugal), Scottish Premier League (Scotland), Liga BBVA (Spain). We aim to predict the outcome of the matches from all leagues in the season 2015/2016 by learning from the data of the remaining seasons. To build our input data vectors and allow for an evaluation in the end, we only consider matches for which the starting lineup and the odds from the betting provider "bet365" are present in our database.

Our first task is to extract suitable features from the given data set. Since the actual lineup of a team has the most impact on the result of the match, we pursue the idea of giving a rating to each player and building our features with this information. To this end, we rely on the "overall player rating" (an integer number between 0 and 100)

given to each player in the "EA Sports FIFA" video game series. These values have been updated on an irregular basis over the last years. Therefore, for each specific soccer match, we always rely on the most recent rating for each player. To build our features, let $h_1, \dots, h_{11}$ be the ratings of the players in the starting lineup of the home team and let $a_1, \dots, a_{11}$ be the corresponding values for the away team. Let, furthermore, the order be such that h_1 and a_1 are the values for the goalkeepers of each team. Then, our features are

1. *team strength*: $\sum_{i=1}^{11} h_i - a_i$,
2. *star bonus*: $\max h_i - \max a_i$,
3. *weakest link*: $\min h_i - \min a_i$,
4. *goalkeeper strength*: $h_1 - a_1$.

The first feature indicates which team has the better total rating, whereas the second one measures how large the gap between the best player of the home team and the best player of the away team is. The third feature is the analogue of the second one for the worst player of each team. Finally, the fourth parameter indicates which goalkeeper is better. Let $T \subset (-1100, 1100) \times (-100, 100)^3$ be the domain in which the parameters reside. Now, the data point $(\mathbf{t}_i, \mathbf{x}_i) \in T \times \mathbb{Z} \subset \mathbb{R}^4 \times \mathbb{R}$, corresponding to the i-th match of our database, consists of a four-dimensional vector $\mathbf{t}_i$, which contains the team strength, the star bonus, the weakest link and the goalkeeper statistics, as well as a label $\mathbf{x}_i \in \mathbb{Z}$. The latter provides the difference between the number of goals scored by the home team and the number of goals scored by the away team, i.e. it is positive if the home team won, zero if the match ended in a draw and negative if the away team won. Our training set, i.e. all points corresponding to matches which took place before the season 2015/2016, is of size $n = 16946$, while the test set on which we evaluate our model consists of 2654 data points.

Learning the data

To measure the quality of our model, we again consider the classification rate. Note, however, that we have three possible outcomes (home team win, draw, away team win) instead of only two as in the previous experiment. Therefore, we introduce a threshold $\delta > 0$ and we classify a match $\mathbf{t}$ as a home team win if $f_{\mathcal{Z}_n,A,\mu}(\mathbf{t}) > \delta$, as an away team win if $f_{\mathcal{Z}_n,A,\mu}(\mathbf{t}) < -\delta$ and as a draw otherwise. Note that, after testing several values for δ, we observed that $\delta = 0$ is the optimal choice in almost every case. Therefore, we set $\delta = 0$ and only classify a match as a home or an away win. Thus, we automatically fail if the match actually ended in a draw.

We run a 5-fold crossvalidation on the training set to determine the optimal choice of $\mu \in \{10^{-1}, 10^{-2}, 10^{-3}, 10^{-4}\}$ and $k_{\text{end}} \in \{2, 3, 4, 5\}$, i.e. the pair of parameters for which the average classification rate is the highest. Subsequently, we learn a model on the full training data set with these optimal parameters and evaluate it on the test data.

Evaluation

With the 5-fold crossvalidation, we found $\lambda = 10^{-2}$ and $k_{\text{end}} = 3$ to be the optimal set of parameters, which achieved an average classification rate of 53.0% on the training data. With these parameters, we obtain a classification rate of 51.1% on the test data. In more detail, we predicted 1034 of 1176 home team wins and 322 of 809 away team wins correctly. Furthermore, there were 669 draws, which we misclassified. The amount of grid points spent in each direction is approximately equal for the coordinates corresponding to the "team strength", the "star bonus" and the "weakest link". In the "goalkeeper strength" direction, however, only a very coarse grid is employed by the dimension-adaptive algorithm. This indicates that more subtle differences in the goalkeepers' abilities do not matter that much for the outcome of a match.

To interpret our quantitative results, note that 44.3% of all matches in the 2015/2016 season ended with a win of the home team. Therefore, we obviously beat the trivial strategy of betting on the home team in each match. Usually, even experts only achieve classification rates around $51 - 53\%$. To this end, let us have a look at the betting odds of "b365", one of the largest betting providers, for all matches in the 2015/2016 season. If we always bet on the outcome of a match according to the smallest betting odds, we achieve a classification rate of 52.0% by classifying 990 home team wins, 391 away team wins and 0 draws correctly. Note that the result of our algorithm comes quite close to this. Note, furthermore, that the computation of the betting odds is usually based on expert knowledge and requires the analysis of many statistics. Therefore, the performance of our method, which is based only on the player ratings from the video games series, is quite remarkable.

Given the b365 odds, we could pose the question if there is any chance to make profit from betting on the matches. With the strategy of betting 1 € on the outcome with the smallest odds for each match, we would actually lose 156.83 €. By betting 1 € on the outcome which the dimension-adaptive sparse grid algorithm predicted, the loss amounts to 118.49 €. A more clever strategy can be obtained by betting on a match $\mathbf{t}$ only if the absolute value $|f_{\mathcal{Z}_n,A,\mu}(\mathbf{t})|$ exceeds a fixed threshold $\omega > 0$. Furthermore, we bet $|f_{\mathcal{Z}_n,A,\mu}(\mathbf{t})|$ € instead of 1 €, i.e. the more confident we are in our prediction, the more money we bet. The profit in dependence on ω can be found in figure 6.21. Note that we are indeed in the positive range for several, large enough choices of ω. The most profit is achieved for $\omega = 1.63$. In this case, we betted 214.99€ to earn a profit of 4.08€. Thus, the ratio between profit and investment is just 1.9%. Furthermore, the profit highly depends on ω and still fluctuates around 0 € also for large choices of ω. Therefore, our approach is only profitable if we guess the right ω by chance.

Overall, we see that the outcome of a soccer match is often influenced by chance/noise and it is impossible to predict it reliably. Nevertheless, we are able to achieve a remarkable classification rate by using statistics from the "EA Sports FIFA" computer game series.

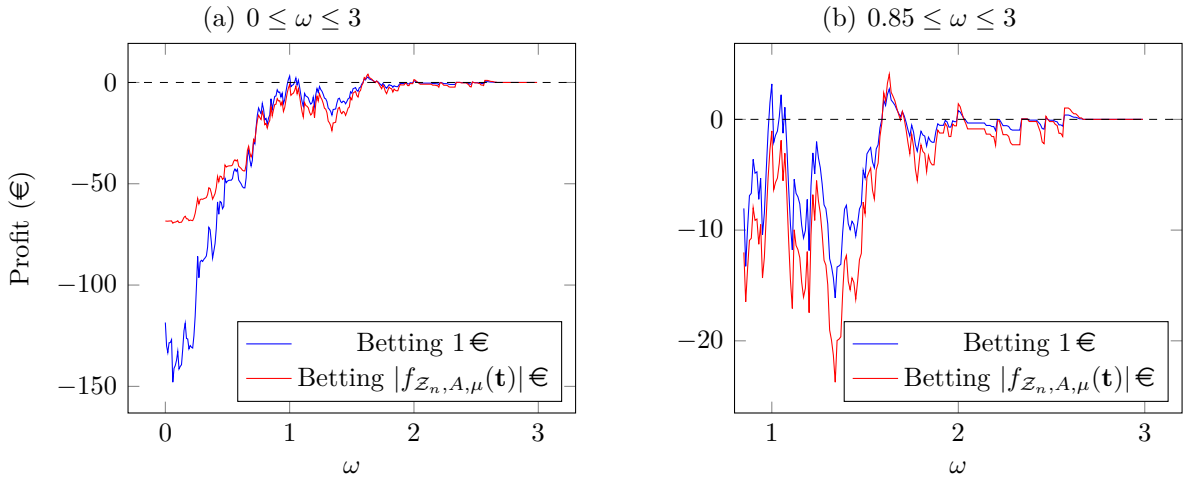

Fig. 6.21: The profit when betting on all soccer matches $\mathbf{t}$ of the season 2015/2016 for which $|f_{\mathcal{Z}_n,A,\mu}(\mathbf{t})| > \omega$. We depicted the results for a 1 € wager and for a $|f_{\mathcal{Z}_n,A,\mu}(\mathbf{t})|$ € wager. The full range ($0 \leq \omega \leq 3$) can be found on the left and a zoomed range ($0.85 \leq \omega \leq 3$) is shown on the right.

7 Conclusion

This chapter serves to provide a short summary of the results of this thesis. Furthermore, we point at open questions and discuss possible directions for future research.

7.1 Summary

In this thesis, we analyzed the vector-valued regression problem over finite-dimensional search spaces. We provided theorems on its solvability/stability and the rate of decay of the overall error for both regularized and unregularized regression. To illustrate our results, we applied them to regression methods based on sparse grids and hyperbolic crosses.

After we had given an overview on the state of the art of regression and had motivated our analysis in the introduction, we provided a short survey on general Lebesgue-Bochner spaces and Sobolev-Bochner spaces and presented a review on the most important properties of vector-valued reproducing kernel Hilbert spaces and real interpolation scales. In the next step, we recapitulated the concepts of spline-based grids and gave a proof on the rate of decay of the L_2 best approximation error for sparse grids. Furthermore, we provided the basic idea behind hyperbolic cross approximation with Fourier polynomials.

We introduced the regression functional and its finite sample counterpart and, subsequently, formulated the general constrained, vector-valued regression problem over arbitrary search sets. We showed that it is solvable if the search set is restricted to a ball in a certain Banach space. Furthermore, we derived sufficient conditions to ensure the uniqueness of the solution. In order to establish upper bounds on the overall error of the regression problem, we split the error into a bias part and a sampling error part and analyzed each part separately. After reviewing known results on the bias for regression over infinite-dimensional search spaces, we explained why these do not apply for the discretization error in finite-dimensional search spaces. Therefore, we proposed a different approach based on Jackson and Bernstein inequalities for this case. Within this framework, we were able to provide a theorem on the decay of the discretization error under the premise that the search set radius is large enough. Then, we proved a probabilistic upper bound on the sampling error in compact finite-dimensional search sets. By applying our results to sparse grid and hyperbolic cross regression, we showed that the corresponding methods can circumvent the curse of dimensionality to some extent, as it is already known from interpolation or approximation problems for instance. Furthermore, we balanced the discretization error and the sampling error to provide

optimal convergence rates for these spaces, which come close to n^{-1}, where n denotes the number of samples. We concluded our analysis of the constrained problem with a comparison to other results in this research area.

In a next step, we related the penalized, Lagrangian dual problem to the constrained one and deduced the system of linear equations which needs to be solved in this case. We also elaborated on the advantages of grid-based methods when solving the dual problem for large n as it is often required in Big Data applications. By extending well-known results for scalar-valued orthonormal basis discretizations in the unregularized case, we were able to prove the stability of both the penalized and the unpenalized approach for arbitrary Riesz basis discretizations. Furthermore, we took a closer look at the special case of noiseless function regression without regularization and established a more refined error analysis here. As before, we applied our theoretical results to sparse grid and hyperbolic cross regression methods and provided sufficient stability conditions and optimally balanced error bounds, which exceed the n^{-1} limit rate from the more general, constrained regression problem. Finally, we related our results to state-of-the-art research.

We validated that our theoretical findings are indeed relevant for practical applications by showing that our upper bounds on the convergence rates come close to the observed rates in numerical experiments. We also introduced a dimension-adaptive variant of the sparse grid regression algorithm, which can be applied to significantly reduce the degrees of freedom which are necessary to deal with anisotropic problems. We concluded with two real-world regression applications.

In summary, we presented a thorough analysis of the stability properties and the error behavior of regularized and unregularized vector-valued regression problems over finite-dimensional search spaces. Our approach is the first one to allow for a detailed analysis of the sparse grid and the hyperbolic cross regression algorithms. The novel results for these examples provide new insights in terms of choosing suitable regularization parameters and determining appropriate couplings between the discretization level and the amount of input data. We confirmed our theoretical findings by several numerical experiments.

7.2 Outlook

There still remain several open questions, which pave the way for complementary studies and future research in the field of regression over finite-dimensional search spaces.

First of all, it is an interesting task to check whether or not the proven upper bounds on the error convergence rates are sharp. As we observed in our numerical experiments, the actual convergence behavior of a sparse grid or hyperbolic cross algorithm often comes very close to our proven bounds. Nevertheless, we have also seen slightly improved rates for specific couplings in the balanced case. Furthermore, it is still an open question, how the optimal coupling and the corresponding convergence rate for hyperbolic cross regression look like if the order of Sobolev smoothness of the solution can be chosen

arbitrarily large.

Although we already derived the optimal rate of convergence which our upper bounds yield when balancing the error terms appropriately, we did not yet take the computational costs into account. This is a sophisticated task by itself since it involves the derivation of a meaningful cost/benefit ratio for the solution of the regression problem with a certain input parameter set as well as the equilibration of the corresponding terms.

Although we saw in our numerical experiments that the dimension-adaptive sparse grid regression method is able to considerably reduce the degrees of freedom which are necessary to deal with an anisotropic problem, there are still open questions from the theoretical point of view: It remains unclear if the regularization parameter has to be adjusted for very anisotropic problems. Furthermore, we did not discuss the optimal coupling between N_k and n and the corresponding convergence rate for a dimension-adaptive sparse grid space.

As we shortly discussed in subsection 4.5.6, there also exist different variations of the sparse grid regression algorithm based on the combination technique or the optimized combination technique. Although our theory is not directly applicable to these variants, it would be interesting to fit our general ideas into this framework. Naturally, the analysis of various other local or global sparse discretizations is interesting by itself.

Finally, an aspect which caught a lot of attention recently is the correct treatment of regression problems over unbounded or infinite-dimensional domains. This problem is often encountered in the field of uncertainty quantification. A first result in this direction can be found in [19] for a polynomial discretization.

Bibliography

[1] R. Adams. *Sobolev Spaces*. Academic Press, New York, 1975.

[2] B. Adcock. Infinite-dimensional ℓ_1 minimization and function approximation from pointwise data. 2016. Preprint available at `http://arxiv.org/abs/1503.02352`.

[3] H. Amann and J. Escher. *Analysis III*. Birkhäuser, 2010.

[4] A. Banerjee, S. Merugu, I. S. Dhillon, and J. Ghosh. Clustering with Bregman divergences. *Journal of Machine Learning Research*, 6:1705–1749, 2005.

[5] R. E. Bellman. *Dynamic Programming*. Princeton University Press, 1957.

[6] I. Bengio and A. Courville. Deep learning. Book in preparation for MIT Press, `http://www.deeplearningbook.org`, 2016.

[7] J. Bergh and J. Löfström. *Interpolation Spaces*. Springer, 1976.

[8] P. Binev, A. Cohen, W. Dahmen, R. DeVore, and V. Temlyakov. Universal algorithms for learning theory - part I: piecewise constant functions. *Journal of Machine Learning Research*, 6:1297–1321, 2005.

[9] P. Binev, A. Cohen, W. Dahmen, R. DeVore, and V. Temlyakov. Universal algorithms for learning theory - part II: piecewise polynomial functions. *Constructive Approximation*, 26:127–152, 2007.

[10] B. Bohn, J. Garcke, and M. Griebel. A sparse grid based method for generative dimensionality reduction of high-dimensional data. *Journal of Computational Physics*, 309:1–17, 2016.

[11] B. Bohn and M. Griebel. An adaptive sparse grid approach for time series predictions. In J. Garcke and M. Griebel, editors, *Sparse grids and applications*, volume 88 of *Lecture Notes in Computational Science and Engineering*, pages 1–30. Springer, 2012.

[12] B. Bohn and M. Griebel. Error estimates for multivariate regression on discretized function spaces. 2014. Submitted to SIAM Journal on Numerical Analysis, also available as INS Preprint No. 1412.

[13] H.-J. Bungartz. *Finite Elements of Higher Order on Sparse Grids*. Habilitation, Department of Informatics, Technical University of Munich, 1998.

[14] H.-J. Bungartz and M. Griebel. Sparse grids. *Acta Numerica*, 13:147–269, 2004.

[15] H.-J. Bungartz, D. Pflüger, and S. Zimmer. Adaptive sparse grid techniques for data mining. 2008.

[16] F. Camastra. Data dimensionality estimation methods: A survey. *Pattern Recognition*, 36(12):2945–2954, 2003.

[17] O. Catoni. *Statistical Learning Theory and Stochastic Optimization*, volume 1851 of *Lecture Notes in Mathematics*. Springer, 2004.

[18] R. Cavoretto, G. Fasshauer, and M. McCourt. An introduction to the Hilbert-Schmidt SVD using iterated Brownian bridge kernels. *Numerical Algorithms*, 68(2):393–422, 2015.

[19] A. Chkifa, A. Cohen, G. Migliorati, F. Nobile, and R. Tempone. Discrete least squares polynomial approximation with random evaluations - application to parametric and stochastic elliptic PDEs. *ESAIM: Mathematical Modelling and Numerical Analysis (M2AN)*, 49(3):815–837, 2015.

[20] A. Chkifa, N. Dexter, H. Tran, and C. Webster. Polynomial approximation via compressed sensing of high dimensional functions on lower sets. Technical Report ORNL/TM-2016/79, Oak Ridge National Laboratory, 2016.

[21] A. Cohen, M. Davenport, and D. Leviatan. On the stability and accuracy of least squares approximations. *Foundations of Computational Mathematics*, 13:819–834, 2013.

[22] F. Cucker and S. Smale. On the mathematical foundations of learning. *Bulletin of the American Mathematical Society*, 39:1–49, 2001.

[23] F. Cucker and D. Zhou. *Learning theory*. Cambridge Monographs on Applied and Computational Mathematics, 2007.

[24] R. DeVore and G. Lorentz. *Constructive Approximation*. A Series of Comprehensive Studies in Mathematics. Springer, 1993.

[25] D. Donoho. For most large underdetermined systems of linear equations the minimal 1-norm solution is also the sparsest solution. *Communications on pure and applied mathematics*, 59:797–829, 2006.

[26] N. Draper and H. Smith. *Applied Regression Analysis*. Wiley, New York, 1998.

[27] D. Dũng and M. Griebel. Hyperbolic cross approximation in infinite dimensions. *Journal of Complexity*, 33:55–88, 2015.

[28] D. Edmunds and H. Triebel. *Function spaces, entropy numbers, differential operators*. Cambridge University Press, 1996.

[29] C. Feuersänger. Dünngitterverfahren für hochdimensionale elliptische partielle Differentialgleichungen. Diploma thesis, Institute for Numerical Simulation, University of Bonn, 2005.

[30] C. Feuersänger. *Sparse Grid Methods for Higher Dimensional Approximation*. PhD thesis, Institute for Numerical Simulation, University of Bonn, 2010.

[31] C. Feuersänger and M. Griebel. Principal manifold learning by sparse grids. *Computing*, 85(4), 2009.

[32] S. Foucart and H. Rauhut. *A Mathematical Introduction to Compressive Sensing*. Applied and Numerical Harmonic Analysis. Birkhäuser, 2013.

[33] A. Galántai. *Projectors and Projection Methods*. Springer, 2004.

[34] F. Galton. Regression towards mediocrity in hereditary stature. *The Journal of the Anthropological Institute of Great Britain and Ireland*, 15:246–263, 1886.

[35] J. Garcke. *Maschinelles Lernen durch Funktionsrekonstruktion mit verallgemeinerten dünnen Gittern*. PhD thesis, Institute for Numerical Simulation, University of Bonn, 2004.

[36] J. Garcke and M. Hegland. Fitting multidimensional data using gradient penalties and the sparse grid combination technique. *Computing*, 84(1-2):1–25, 2009.

[37] C. Gauß. *Theoria Motus Corporum Coelestium in Sectionibus Conicis Solem Ambientium*. Frid. Perthes & I.H. Besser, Hamburg, 1809.

[38] M. Griebel. Adaptive sparse grid multilevel methods for elliptic PDEs based on finite differences. *Computing*, 61(2):151–179, 1998.

[39] M. Griebel and H. Harbrecht. On the construction of sparse tensor product spaces. *Mathematics of Computations*, 82(282):975–994, 2013.

[40] M. Griebel and M. Hegland. A finite element method for density estimation with Gaussian priors. *SIAM Journal on Numerical Analysis*, 47(6):4759–4792, 2010.

[41] M. Griebel and P. Oswald. Tensor product type subspace splitting and multilevel iterative methods for anisotropic problems. *Advances in Computational Mathematics*, 4:171–206, 1995.

[42] M. Griebel, P. Oswald, and T. Schiekofer. Sparse grids for boundary integral equations. *Numerische Mathematik*, 83(2):279–312, 1999.

[43] M. Griebel, C. Rieger, and B. Zwicknagl. Multiscale approximation and reproducing kernel Hilbert space methods. *SIAM Journal on Numerical Analysis*, 53(2):852–873, 2015.

[44] M. Griebel, C. Rieger, and B. Zwicknagl. Regularized kernel based reconstruction in generalized Besov spaces. *Accepted by Foundations of Computational Mathematics*, 2016.

[45] S. Gudder and D. Strawther. Strictly convex normed linear spaces. *Proceedings of the American Mathematical Society*, 59(2):263–267, 1976.

[46] L. Györfi, M. Kohler, A. Krzyżak, and H. Walk. *A Distribution-Free Theory of Nonparametric Regression*. Springer, 2002.

[47] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer, 2001.

[48] M. Hegland. Data mining techniques. *Acta Numerica*, 10:313–355, 2001.

[49] H. Kantz and T. Schreiber. *Nonlinear Time Series Analysis*. Cambridge University Press, 2004.

[50] R. Katayama, K. Kuwata, Y. Kajitani, and M. Watanabe. Embedding dimension estimation of chaotic time series using self-generating radial basis function network. *Fuzzy Sets and Systems*, 71:311–327, 1995.

[51] J. Kelley. *General Topology*. Springer, 1991.

[52] S. Knapek. *Approximation und Kompression mit Tensorprodukt-Multiskalenräumen*. PhD thesis, Institute for Numerical Simulation, University of Bonn, 2000.

[53] M. Kohler. Inequalities for uniform deviations of averages from expectations with applications to nonparametric regression. *Journal of Statistical Planning and Inference*, 89(1-2):1–23, 2000.

[54] S. Konyagin and V. Temlyakov. The entropy in learning theory. Error estimates. *Constructive Approximation*, 25:1–27, 2007.

[55] J. Lee and M. Verleysen. *Nonlinear Dimensionality Reduction*. Springer Science, 2007.

[56] A. Legendre. *Nouvelles méthodes pour la détermination des orbites des comètes*. Firmin Didot, Paris, 1805.

[57] M. Lichman. UCI machine learning repository, 2013. `http://archive.ics.uci.edu/ml`, University of California, Irvine, School of Information and Computer Sciences.

[58] C. Micchelli and M. Pontil. On learning vector-valued functions. *Neural Computation*, 17:177–204, 2005.

[59] G. Migliorati and F. Nobile. Analysis of discrete least squares on multivariate polynomial spaces with evaluations at low-discrepancy point sets. *Journal of Complexity*, 31(4):517–542, 2015.

[60] G. Migliorati, F. Nobile, and R. Tempone. Convergence estimates in probability and in expectation for discrete least squares with noisy evaluations at random points. *Journal of Multivariate Analysis*, 142:167–182, 2015.

[61] G. Migliorati, F. Nobile, E. von Schwerin, and R. Tempone. Analysis of discrete L^2 projection on polynomial spaces with random evaluations. *Foundations of Computational Mathematics*, 14:419–456, 2014.

[62] K. Miller. Complex linear least squares. *SIAM Review*, 15(4):706–726, 1973.

[63] H. Minh, L. Bazzani, and V. Murino. A unifying framework for vector-valued manifold regularization and multi-view learning. In S. Dasgupta and D. McAllester, editors, *Proceedings of the 30th International Conference on Machine Learning (ICML-13)*, volume 28, pages 100–108. JMLR Workshop and Conference Proceedings, 2013.

[64] T. Muramatu. Besov spaces and Sobolev spaces of generalized functions defined on a general region. *Publications of the Research Institute for Mathematical Sciences, Kyoto University*, 9:325–396, 1974.

[65] J. Oettershagen. Reduktion der effektiven Dimension und ihre Anwendung auf hochdimensionale Probleme. Diploma thesis, Institute for Numerical Simulation, University of Bonn, 2011.

[66] V. Paulsen and M. Raghupathi. *An Introduction to the Theory of Reproducing Kernel Hilbert Spaces*. Cambridge University Press, 2016.

[67] J. Peetre. *A theory of interpolation of normed spaces*, volume 39 of *Notas de Matemática*. Rio de Janeiro: Instituto de Matemática Pura e Aplicada, Conselho Nacional de Pesquisas, 1968.

[68] D. Pflüger, B. Peherstorfer, and H.-J. Bungartz. Spatially adaptive sparse grids for high-dimensional data-driven problems. *Journal of Complexity*, 26(5):508–522, 2010.

[69] T. Poggio, R. Rifkin, S. Mukherjee, and P. Niyogi. General conditions for predictivity in learning theory. *Nature*, 428:419–422, 2004.

[70] W. Press, S. Teukolsky, W. Vetterling, and B. Flannery. *Numerical Recipes: The Art of Scientific Computing (Third Edition)*. Cambridge University Press, 2007.

[71] O. Roesler and D. Suendermann. A first step towards eye state prediction using EEG. In *Proceedings of the International Conference on Applied Informatics for Health and Life Sciences*, 2013. Instanbul, Turkey.

[72] B. Scharf, H.-J. Schmeißer, and W. Sickel. Traces of vector-valued Sobolev spaces. *Mathematische Nachrichten*, 285(8-9):1082–1106, 2012.

[73] H.-J. Schmeisser. Recent developments in the theory of function spaces with dominating mixed smoothness. In J. Rákosník, editor, *Proceedings on Nonlinear Analysis, Function Spaces and Applications, Institute of Mathematics of the Czech Academy of Sciences, Prague*, volume 8, pages 145–204, 2007.

[74] B. Schölkopf and A. Smola. *Learning with Kernels – Support Vector Machines, Regularization, Optimization, and Beyond*. The MIT Press – Cambridge, Massachusetts, 2002.

[75] S. Smale and D. Zhou. Estimating the approximation error in learning theory. *Analysis and Applications*, 1(1):17–41, 2003.

[76] G. Sparr. Interpolation of several banach spaces. *Annali di Matematica Pura ed Applicata*, 99(1):247–316, 1974.

[77] M. Steffens, T. Becker, T. Sander, R. Fimmers, C. Herold, D. Holler, C. Leu, S. Herms, S. Cichon, B. Bohn, T. Gerstner, M. Griebel, M. Nöthen, T. Wienker, and M. Baur. Feasible and successful: Genome-wide Interaction Analysis (GWIA) involving all 1.9×10^{11} pair-wise interaction tests. *Human Heredity*, 69:268–284, 2010.

[78] V. Temlyakov. *Approximation of Periodic Functions*. Nova Science, 1993.

[79] V. Temlyakov. Approximation in learning theory. *Constructive Approximation*, 27:33–74, 2008.

[80] H. Triebel. *Interpolation Theory, Function Spaces, Differential Operators*. North Holland Mathematical Library, 1978.

[81] H. Triebel. *Fractals and Spectra*, volume 91 of *Monographs in Mathematics*. Birkhäuser, 1997.

[82] H. Triebel. *Bases in Function Spaces, Sampling, Discrepancy, Numerical Integration*, volume 11 of *Tracts in Mathematics*. European Mathematical Society, 2010.

[83] J. Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of Computational Mathematics*, 12(4):389–434, 2011.

[84] T. Ullrich. *Smolyak's Algorithm, Sparse Grid Approximation and Periodic Function Spaces with Dominating Mixed Smoothness*. PhD thesis, University of Jena, 2007.

[85] V. Vapnik. *Statistical Learning Theory*. John Wiley & Sons, 1998.

[86] J. Vybiral. *Function spaces with dominating mixed smoothness*. PhD thesis, University of Jena, 2005.

[87] I. Witten and E. Frank. *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2005.

[88] M. Wong and M. Hegland. Maximum a posteriori density estimation and the sparse grid combination technique. In S. McCue, T. Moroney, D. Mallet, and J. Bunder, editors, *Proceedings of the 16th Biennial Computational Techniques and Applications Conference, CTAC-2012*, volume 54, pages 508–522, 2013.

[89] Y. Ying and D. Zhou. Learnability of Gaussians with flexible variances. *Journal of Machine Learning Research*, 8:249–276, 2007.

[90] E. Zeidler. *Nonlinear Functional Analysis and its Applications III - Variational Methods and Optimization*. Springer, New York, 1985.

[91] A. Zeiser. Fast matrix-vector multiplication in the sparse-grid galerkin method. *Journal of Scientific Computing*, 47(3):328–346, 2011.

[92] Y. Zhang, F. Cao, and Z. Xu. Estimation of learning rate of least square algorithm via Jackson operator. *Neurocomputing*, 74:516–521, 2011.